Discussion Paper Series – CRC TR 224

Discussion Paper No. 149
Project B 07

# Complementarities in Behavioral Interventions: Evidence from a Field Experiment on Energy Conservation

Ximeng Fang*
Lorenz Goette**
Bettina Rockenbach***
Matthias Sutter****
Verena Tiefenbeck†
Samuel Schoeb††
Thorsten Staake†††

August 2021
*(First version: January 2020)*

*University of Bonn (email: x.fang@uni-bonn.de)
**University of Bonn, National University of Singapore
***University of Cologne
****Max Planck Institute for Research into Collective Goods, University of Cologne, University of Innsbruck
†Friedrich-Alexander Universität Nürnberg-Erlangen, ETH Zurich
††University of Bamberg
†††University of Bamberg

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

# Complementarities in Behavioral Interventions: Evidence from a Field Experiment on Energy Conservation*

Ximeng Fang Lorenz Goette Bettina Rockenbach Matthias Sutter
Verena Tiefenbeck Samuel Schoeb Thorsten Staake ‡

*30 August 2021*

**Abstract**

Behavioral policy often aims at overcoming biases due to, e.g., imperfect information or inattention. When there are multiple sources of bias, interventions targeting different sources each may be complements: each intervention becomes more effective when combined with others. We test this in a field experiment on energy conservation in a resource-intensive everyday activity (showering). One intervention, shower energy reports, primarily improves knowledge about environmental impacts; another intervention, real-time feedback, primarily increases salience of resource use. While only the latter reduced energy consumption when implemented in isolation, combining both interventions boosted this conservation effect by over 50%, indicating a striking complementarity.

*JEL classification:* D12, D83, Q41

*Keywords:* behavioral interventions, energy conservation, inattention, information provision, policy interactions, randomized controlled trials, real-time feedback, salience of information

[*] We are indebted to our research assistant team (in particular Benedikt Kauf and Kristina Steinbrecher) for their indefatigable support in running the field experiment, and to Lim Zhi Hao for editing. We would further like to thank audiences in Bonn, at the EEA Virtual 2020, the ECONtribute Retreat 2020, the ESA Dijon, the IAREP-SABE Dublin, the EAERE Manchester, the 3rd CRC TR 224 Conference, the PCBS 2019 Prague, the 12th RGS Doctoral Conference, and the 2nd BoMa PhD Workshop for helpful comments. Financial support by the University of Cologne Forum "Energy", the SUN Institute Environment & Sustainability, and by the Deutsche Forschungsgemeinschaft (DFG) through CRC TR 224 (Project B07) is gratefully acknowledged. The 2019 supplementary survey was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2126/1– 390838866. This study is registered in the AEA RCT Registry under the identification number "AEARCTR-0004865".

[‡] Fang: University of Bonn (email: x.fang@uni-bonn.de). Goette: University of Bonn, National University of Singapore. Rockenbach: University of Cologne. Sutter: Max Planck Institute for Research into Collective Goods, University of Cologne, University of Innsbruck. Tiefenbeck: Friedrich-Alexander Universität Nürnberg-Erlangen, ETH Zurich. Schoeb: University of Bamberg. Staake: University of Bamberg, ETH Zurich.

# 1. Introduction

Amidst growing public concern about climate change and resource scarcity, many individuals intend to make personal sacrifices to protect the environment; yet they often fail to act pro-environmentally in their everyday lives (Kollmuss and Agyeman, 2002; Frederiks, Stenner and Hobman, 2015). This gap between intentions and actions can result from a multiplicity of behavioral frictions and biases. For instance, previous research has shown that individuals tend to underestimate the impact of highly resource-intensive behaviors (Attari et al., 2010; Attari, 2014), and that they may also not be fully attentive to their resource use (Allcott, 2016; Tiefenbeck et al., 2018).

Other factors such as self-control problems and status quo bias may certainly also play a role. Importantly, however, such behavioral biases could not only prevent consumers from acting on their intrinsic prosocial or pro-environmental motivations, but also mute their response to policy interventions aimed at encouraging behavioral change. Thus, when multiple dimensions of bias are present at the same time, interventions that miss an important dimension may fail to unfold their full potential. For example, providing information about environmental impacts may have little effect on behavior if individuals remain inattentive to their resource use.[1] Conversely, making resource use salient may only have a muted effect if agents remain unaware of adverse environmental impacts. Hence, in this example, a combined approach that targets both imperfect information and inattention could have synergetic, mutually reinforcing effects, i.e. positive interaction effects or complementarities. More generally, we argue that bundling interventions can result in complementarities if each intervention is particularly suited to address a different source of behavioral bias. Following Coe and Snower (1997), we define interventions as complements if *each* intervention becomes more effective when implemented in conjunction with the other(s) than in isolation. While many studies consider the use of combined interventions, there is need for more theoretical and empirical research that investigates systematic drivers of complementarity (or substitutability) and thereby provides guidance for the design of effective behavioral policy.

In this paper, we report evidence from a three-month randomized field experiment in which we used two well-studied behavioral policy tools to encourage resource conservation in an energy- and water-intensive everyday activity, namely showering. Our interventions were designed in such a way that we target different potential sources of behavioral bias against resource conservation. The first intervention, shower energy reports, inspired by the Opower home energy reports (Allcott, 2011), were primarily aimed at closing knowledge gaps about environmental impacts by providing information on water use as well as on energy use and $CO_2$ emissions due to water heating. The second intervention, real-time feedback, provided immediately visible and salient information

[1] Information provision is often regarded as a promising policy lever, as individuals often misperceive the environmental impact of everyday activities (Attari et al., 2010; Attari, 2014; Camilleri et al., 2019) and tend to engage in relatively ineffective conservation measures (Gardner and Stern, 2008; Tonke, 2019).

on water consumption — but not energy use or CO2 emissions — through a smart meter display (Tiefenbeck et al., 2018), and could thus help individuals focus their attention while they engaged in the activity. Crucially, we implemented a complete 2×2-design to evaluate both the combined intervention as well as each intervention in isolation, which allows us to uncover potential complementarities.

To formalize our argument as to why complementarities might arise in such a context, we introduce a stylized theoretical framework in which biased perceptions of resource use arise from multiple sources (e.g. imperfect information, limited attention). Each of these biases acts akin to a discount factor and thus prevents agents from fully incorporating the marginal costs of resource consumption into their behavior. A key prediction from our framework is that when each bias mutes the perceived cost of resource use independently of other biases, then the effects of pro-environmental interventions that mitigate different sets of biases can reinforce each other, so that the interventions become complements. The intuition is simple: the more one particular bias is reduced, the larger is the impact of reducing another bias. For example, the more attention an agent pays to her resource use behavior, the more likely it is that she will actually change her behavior when learning that the environmental impact is more negative than previously thought. This interaction mechanism is absent when two interventions mostly operate through the same behavioral channel, e.g. if they both provide the same type of information.

There are several reasons why (warm) water consumption in the shower provides an interesting context for studying complementarities in behavioral interventions. First, showering is a resource-intensive activity: an average shower in our sample requires 2.2 kWh of energy to heat up 38 liters of water, which corresponds to about 10% of the average residential energy use and 30% of the average water consumption per capita and day in Germany, where we conducted our study.[2] Second, individuals tend to underestimate the $CO_2$ emissions caused by warm water consumption in the shower — by as much as 89% on average based on one of our surveys —, which creates scope for reducing energy consumption through information provision. Third, showering is also prone to behavioral biases like limited attention and self-control problems, as the pleasure of a warm shower is salient and immediate, whereas the cost of resource use seems abstract and is hard to keep track of (Tiefenbeck et al., 2018). Since individuals may not fully engage in conservation efforts unless they are informed about the actual impact of their behavior *and* keep environmental concerns on top of their minds while showering, it may be necessary to draw on both of these mechanisms at the same time.

We conducted our field experiment in student dormitories in the cities of Bonn and Cologne, Germany, in the winter term 2016/17. A total of 351 students participated in our experiment, with all of them living in single-person dorm apartments with a private bathroom. For the duration of our study, from early December 2016 until early March

[2] Calculated based on information from the German Federal Statistical Office. Source: `https://www.destatis.de/EN/Themes/Society-Environment/Environment/_node.html`

2017, each participant was equipped with a smart shower meter (installed directly below the shower head) that recorded detailed data of each water extraction. Subjects were randomly assigned into one of four experimental conditions: no intervention (CON group), shower energy reports only (SER group), real-time feedback only (RTF group), or both interventions combined (DUAL group). After a baseline stage of 10 showers, the smart meter started displaying real-time feedback on water use for subjects in RTF and DUAL. About halfway into the study, we further started constructing the individualized energy reports using uploaded data from the smart meters and sent them out to subjects in SER and DUAL via email. This staggered design allows us to identify and estimate treatment effects of each intervention regime in a difference-in-differences setup. The shower energy reports mainly aimed at reducing knowledge gaps about environmental impacts, whereas real-time feedback mainly aimed at focusing attention and creating a sense of immediacy. As both mechanisms might be important for encouraging conservation behavior, we hypothesize that the two interventions are complements.[3]

Our empirical results show that, compared to the control group, subjects in the RTF group reduced their energy (water) consumption by about 0.4 kWh (6.3 liters) per shower, which corresponds to 17–18% of baseline resource use. This treatment effect remains stable over the entire 3-month duration of the study. Energy reports in isolation (SER group) did not lead to any statistically detectable conservation effects. However, in line with our hypothesis, we observe a striking complementarity between the two interventions. Combining energy reports with real-time feedback (DUAL group) *further* increased the treatment effect of real-time feedback in isolation by 0.22 kWh of energy (3.8 liters of water) per shower, i.e. by more than 50%. Hence, the shower energy reports simply appeared to require an enhanced choice environment to become effective. The additional reduction of energy use in the DUAL group was not driven by short-lived boosts directly after receiving a shower energy report, but rather seemed to unfold over time, which speaks against Hawthorne or pure reminder effects as the underlying mechanism. Furthermore, we generally find no evidence of adjustments on the extensive margin, i.e. the number of showers people take. One noteworthy feature of our sample is that subjects had no monetary incentives for conserving energy or water, since they paid a flat fee for utilities. Thus, all conservation effects are driven solely by non-monetary motives, which makes them even more remarkable.

Additional questionnaire data shows that both interventions helped subjects form more precise beliefs about their own water use in the shower; there is no evidence that subjects in the DUAL group read their reports more carefully than subjects in the SER group. Supplementary survey results from a comparable sample further suggest that informa-

[3] Complementarity can also arise if our interventions do not exactly work through the described mechanisms, as long as they sufficiently differ from each other in their targeted sources. For example, real-time feedback could be interpreted as information provision about instantaneous water consumption that can facilitate learning or optimization, and this information can be complementary to the information on $CO_2$ emissions provided through shower energy reports.

tion included in shower energy reports also induces drastic (upward) updates in beliefs about $CO_2$ emissions due to warm water consumption in the shower. Hence, the null result for shower energy reports in isolation is not due to lack of learning. Instead, it seems that in the absence of real-time feedback, inattention and lack of immediate visibility have prevented knowledge gains about environmental impacts from translating into effective conservation behavior.

Overall, our findings are consistent with the hypothesis that the presence of multiple bias dimensions can induce complementarities between interventions that largely operate through different behavioral mechanisms. This implies that appropriate policy bundling may increase the cost-effectiveness of interventions beyond what can be achieved with piecemeal approaches. In particular, lack of evidence for effectiveness of an intervention in isolation — as for information provision through shower energy report in our case — does not imply that it cannot be effective in an enhanced policy environment that also takes into consideration further potential sources of bias.

Our study builds on important previous contributions that have studied the effects of similar behavioral interventions on household energy conervation.[4] For example, in an influential evaluation of the Opower home energy reports, which provide information on aggregate electricity use to millions of U.S. households, Allcott (2011) reports a household-level conservation effect of 2%, or about 0.62 kWh per day; effectivity might be smaller outside the U.S., where the baseline energy consumption tends to be lower (see e.g. Andor et al., 2020, for a sample of German households), or when there are little monetary incentives to save energy (Myers and Souza, 2019). Our SER intervention also gives feedback about past consumption patterns, although differing to classical home energy reports in several aspects, mainly in that it targets one specific activity (showering) instead of aggregate household consumption. Disaggregated, activity-specific feedback could enable better learning and thus stronger conservation responses in the targeted activities (Gerster, Andor and Goette, 2020), in particular when provided in shorter time intervals or even in real time. Tiefenbeck et al. (2018) provide real-time feedback in the shower through the same type of smart meter that we use in this study and document a 22% conservation effect, or, in absolute terms, a reduction of 0.6 kWh energy and 9 liters of water per shower. These results also replicate in a sample without monetary incentives and without self-selection into the study (Tiefenbeck et al., 2019). As real-time feedback can make resource consumption immediately salient, a natural question is whether we can use this to improve the effectiveness of other interventions that aim to encourage conservation behavior through further mechanisms like more detailed information provision or social norms and could thus benefit from generally higher attention to pro-environmental motives.

[4] Pro-environmental interventions have drawn from a broad set of instruments such as information provision, social norms, goal-setting, etc. For reviews, see e.g. Abrahamse et al. (2005), Fischer (2008), Delmas, Fischlein and Asensio (2013), Karlin, Zinger and Ford (2015), Andor and Fels (2018), Carlsson et al. (2021).

We further relate to a number of other studies that test a combination of different interventions, especially to studies on pro-environmental behavior that also consider the idea that policy measures might become more effective when implemented in conjunction with others.[5] For example, Jessoe and Rapson (2014) find that pricing schemes that incentivize lower peak electricity consumption can fail to change behavior due to consumers not knowing how to effectively adjust electricity usage; only households who have been outfitted with in-home-displays reduce consumption significantly in response to price hikes. Other recent studies who investigate the combination of financial incentives and behavioral interventions tend to find that they affect behavior along different margins or for different subpopulations, but find no conclusive patterns with regard to interaction effects (List et al., 2017; Holladay et al., 2019; Giaccherini et al., 2020; Fanghella, Ploner and Tavoni, 2021). Hahn et al. (2016) test the individual and combined effects of social comparisons and loss framing on take-up of water-efficient technology as well as general household water consumption, but the results for interaction effects are mixed. Brandon et al. (2019) evaluate the interaction effect of two behavioral interventions on household energy conservation, home energy reports and "peak energy reports", which provide feedback and social norms for households' peak electricity use. As both interventions are very similar and likely operate through similar behavioral channels, it is not clear whether one should expect any interaction effect. Indeed, Brandon et al. find neither strong evidence for complementarity nor substitutability. While we add to this literature by providing a novel case study on the complementarity of two specific types of behavioral interventions, our main contribution is that we attempt to make a step towards understanding mechanisms that systematically lead different policy interventions to become complements or substitutes. Hence, the empirical design is embedded within a conceptual framework — highlighting specifically the role of multiple sources of behavioral bias — that can be adapted to form hypotheses about policy interactions in other contexts as well.

The remainder of this paper is structured as follows: Section 2 introduces the theoretical framework for policy interactions under multiple sources of behavioral bias. Section 3 describes the experimental setup and derives behavioral predictions. Section 4 presents our data as well as some descriptive statistics. Section 5 explains our empirical approach

[5] Combined interventions are also used in other contexts than pro-environmental behavior. For example, in development economics, a number of studies experimentally test the combined effect of different interventions on financial savings (Dupas and Robinson, 2013; Jamison, Karlan and Zinman, 2014), education (Mbiti et al., 2019), risky sexual behavior (Duflo, Dupas and Kremer, 2015; Dupas, Huillery and Seban, 2018), demand for health products (Ashraf, Jack and Kamenica, 2013), or immunization (Banerjee et al., 2021). Many of these studies, however, cannot explicitly test policy interactions, and none of them asks more generally if or why different interventions can be complements if they target separate mechanisms. One notable study is by Mbiti et al. (2019), who find complementarities between providing school grants and adding teacher incentives in improving children's educational outcomes. Another study by Banerjee et al. (2021) employs reminders, incentives, and information ambassador interventions on a large-scale, and then uses a data-driven approach to identify the best combination; in particular, one observation is that information ambassadors seem to amplify the effect of other interventions.

and Section 6 presents our main empirical results. In Section 7, we study the potential mechanisms underlying the results and provide robustness checks. Section 8 concludes.

## 2. Theoretical framework

We begin by introducing a stylized framework to formalize our argument of how complementarities in behavioral interventions can arise in settings with multiple sources of biased perceptions, e.g. imperfect information, limited attention, present bias.

### 2.1. Setup

*Basic setup.* — The agent engages in an energy-intensive activity, say showering and the policy objective is to reduce energy use. Her consumption level is determined by a trade-off between the consumption utility (incl. pleasure, instrumental benefits, opportunity costs of time) and the perceived costs of resource use (incl. monetary costs, environmental concern). She chooses energy use level $e \geq 0$ to maximize

$$U(e) = V(e) - B \cdot C(e)\,, \tag{1}$$

where $V(e)$ is the instantaneous consumption utility and $C(e)$ is the cost of energy consumption.[6] In addition to standard smoothness conditions, we assume that $V$ is hump-shaped (locally increasing at 0, strictly concave, unique maximum) and that $C$ is strictly monotonically increasing and weakly convex. For simplicity, we abstract from uncertainty or dynamics. In the absence of monetary motives, as in our empirical setting, $C(e)$ is the "moral" cost the agent perceives in face of the negative externalities from energy use. However, the cost function is attenuated by an aggregate bias factor $B$, and energy use is biased upwards if $B \in [0, 1)$.

*Multiple sources of bias.* — The aggregate $B$ factor can be the product of a collection of separate factors. To illustrate the mechanics, it is sufficient to focus on the simple case with two sources of bias:

$$B = b_1 \cdot b_2\,. \tag{2}$$

For example, the first factor $b_1$ may indicate the degree to which the agent underestimates energy intensity (as shown, e.g., in Attari et al., 2010), and the second factor $b_2$ the degree to which she is inattentive (e.g., Tiefenbeck et al., 2018). The multiplicative form captures that any single factor can independently prevent the agent from implementing her conservation motive. In this example, the agent will not take into account environmental cost both if she believes her behavior has no impact ($b_1 = 0$) *and* if she is fully inattentive

[6] The agent may not explicitly optimize with regard to energy use, but as long as the mapping from actual decision variable (e.g. shower duration) to resource use is injective, we can represent the problem *as if* the agent was optimizing over energy use.

($b_2 = 0$), either condition by itself is sufficient.[7] Note that the entire framework can be easily generalized to the case of $B = \prod_{k=1}^{K} b_k$ with $k > 2$.

*Consumption behavior.* — The agent's consumption choice is defined by the intersection of marginal utility and marginal costs, but with the latter being diminished by the aggregate bias:

$$V'(e) \; = \; B \cdot C'(e)\,. \tag{3}$$

If $B < 1$, then the marginal cost is underweighted and energy use is thus biased upwards. Defining $f$ such that $f(e) = \frac{V'(e)}{C'(e)}$ for all $e \in [0, \infty)$, we can directly map the relation between implemented energy use and aggregate bias as

$$e(B) \; = \; f^{-1}(B)\,, \tag{4}$$

because equation (3) implies that $f(e(B)) = B$. Notice that $f^{-1}$ is a strictly decreasing function, so the weaker the aggregate bias, i.e. $B$ closer to 1, the lower the energy use.[8] In this sense, $B$ can be interpreted as an input for energy conservation.

*Behavioral interventions.* — In this setup, we define behavioral interventions as policies that aim to change consumers' behavior by changing $B$.[9] In contrast, price-based policies would be aimed at increasing the marginal costs of energy use, $C'(e)$, that the agent faces.[10] As $B = b_1 \cdot b_2$, there are two behavioral policy levers for reducing energy consumption: raising $b_1$ (e.g. providing information) and raising $b_2$ (e.g. enhancing salience).

## 2.2. Policy interaction mechanisms

Two interventions, $X$ and $Y$, are complements if their combination reduces bias by more than the sum of their individual effects, i.e $\Delta B^{X+Y} > \Delta B^{X} + \Delta B^{Y}$. If they are substitutes, the inequality sign is reversed. Notice that even under substitutability, it can be the case that $X + Y$ is more effective than either $X$ or $Y$ in isolation, i.e. $\Delta B^{X+Y} > \Delta B^{X}$ and $\Delta B^{X+Y} > \Delta B^{Y}$. Thus, to empirically identify interaction effects between different

[7] This is reminiscent of the Anna Karenina principle, which states that failure in a single factor may lead to failure of an endeavor as a whole. It is inspired by the opening phrase of Leo Tolstoy's novel *Anna Karenina*: "Happy families are all alike; every unhappy family is unhappy in its own way." (Tolstoy, 2003).

[8] This is because marginal consumption utility $V'(e)$ is strictly decreasing and marginal cost $C'(e)$ is non-decreasing. Hence, $f$ is strictly increasing, so the inverse function $f^{-1}$ exists and is strictly decreasing.

[9] Equation (4) shows that any policy $X$ that mitigates the aggregate bias ($B^X$) compared to no-intervention state $B$ will induce the agent to conserve energy. Hence, $\Delta B^X = B^X - B > 0$ implies that $\Delta e^X = e(B^X) - e(B) < 0$. The more successful an intervention is in mitigating the aggregate bias, the larger the energy reduction effect.

[10] Our framework also allows for an interpretation that takes more a social planner's point of view, aiming for the agent to internalize the full social cost $C^s(e)$. The ratio of private to social cost $C(e)/C^s(e)$ would then be another factor entering into the aggregate bias $B^s$, so decision utility is $U(e) = V(e) - B^s \cdot C^s(e)$. This interpretation highlights the overarching policy objective of reducing externalities instead of "internalities". Efforts to increase the privately perceived cost can include Pigouvian taxes (e.g. carbon pricing), social norms, goal-setting, etc.

policy interventions, it is also necessary to evaluate the effectiveness of each intervention in isolation. Our theoretical framework allows for several mechanisms that could make interventions either complements or substitutes.

*Complementary policy levers.* — The key mechanism we aim to highlight in this paper is that in the presence of multiple sources of bias, policies that target only one dimension may have a limited effect on behavior, whereas the effect of combining several policy levers may be superadditive. This is an immediate consequence of the multiplicative structure of $B$, which implies a positive cross-derivative: $(\partial^2 B / \partial b_1 \partial b_2) > 0$. For example, correcting perceptions of the environmental impact $b_1$ may only have a small impact on behavior if the attention parameter $b_2$ is still close to zero.

There is a simple geometric interpretation to illustrate this: the overall bias parameter $B$, defined in equation (2), can be thought of as the area of a rectangle with sides of lengths $b_1$ and $b_2$ (see Figure 1a). The larger the rectangle the lower the resulting energy consumption will be. Now suppose that $b_1$ is exogenously increased by $\delta_1$. The resulting increase in $B$ will be $\delta_1 b_2$, as it is attenuated by $b_2$. Analogously, an exogenous increase of $\delta_2$ in the dimension of $b_2$ results in an aggregate change of $\delta_2 b_1$. The effect of jointly increasing $b_1$ and $b_2$ by the same amounts, however, results in an overall change of

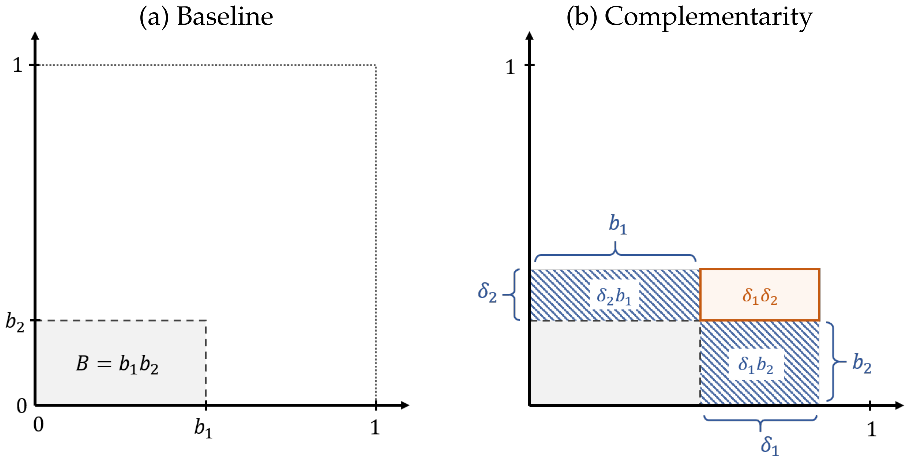
$$\Delta B \ = \ \delta_1 b_2 \ + \ \delta_2 b_1 \ + \ \delta_1 \delta_2 \,. \tag{5}$$

There is an additional effect of size $\delta_1 \delta_2$, because a gain in one dimension also makes the improvement in the other dimension larger. Geometrically, this is represented by the top right rectangle outlined in the second graph of Figure 1. This mechanism potentially induces complementarity between interventions that are specialized on mitigating different sources of bias each.

In practice, it may be hard to design "pure" interventions where each intervention changes only one dimension of $B$. To illustrate the complementarity in an example that might be closer to reality, consider the case of two sources of bias and two interventions, $X$ and $Y$. Suppose that intervention $X$ is primarily targeted at the perception of the environmental impact $b_1$, while potentially also having a positive side-effect on $b_2$, which could describe an information intervention which may also lead to endogenously higher attention levels (Hanna, Mullainathan and Schwartzstein, 2014; Gabaix, 2017). Analogously, intervention $Y$ is primarily targeted at the attention parameter $b_2$, with positive side-effects on $b_1$. This could describe a salience intervention that incindentally also offers some degree of information or induces information search efforts. Hence, the relevant parameters are such that $\delta_1^X \geq \delta_1^Y$ and $\delta_2^Y \geq \delta_2^X$. The reduction in bias of each intervention in isolation are $\Delta B^X = \delta_1^X b_2 + \delta_2^X b_1 + \delta_1^X \delta_2^X$ and $\Delta B^Y = \delta_1^Y b_2 + \delta_2^Y b_1 + \delta_1^Y \delta_2^Y$, respectively, which is also illustrated in Figure 2a and b.

*Aggregating policy interventions.* — When two partially overlapping interventions are

Figure 1: Depiction of example interventions



*Notes.* The grey rectangle in Figure (a) illustrates the aggregate bias $B$ as defined in equation 2 without any intervention in place. Figure (b) illustrates the increase in $B$ through exogenous interventions in each dimension.

introduced jointly, we need to specify how they aggregate into the overall bias $B$. As a benchmark, we assume that the mitigation effects $\delta_i^X, \delta_i^Y$ are additive (and that the resulting $b_i$ does not exceed 1). Figure 2c illustrates this example, in which $\delta_1^{X+Y} = \delta_1^X + \delta_1^Y$ and $\delta_2^{X+Y} = \delta_2^X + \delta_2^Y$. The additional bias reduction is

$$\Delta B^{X+Y} - \Delta B^X - \Delta B^Y = \delta_1^X \delta_2^Y + \delta_2^X \delta_1^Y. \tag{6}$$

Notice, that — holding constant $\delta_1^{X+Y}$ and $\delta_2^{X+Y}$ — the potential for complementarity is largest for two completely specialized interventions.

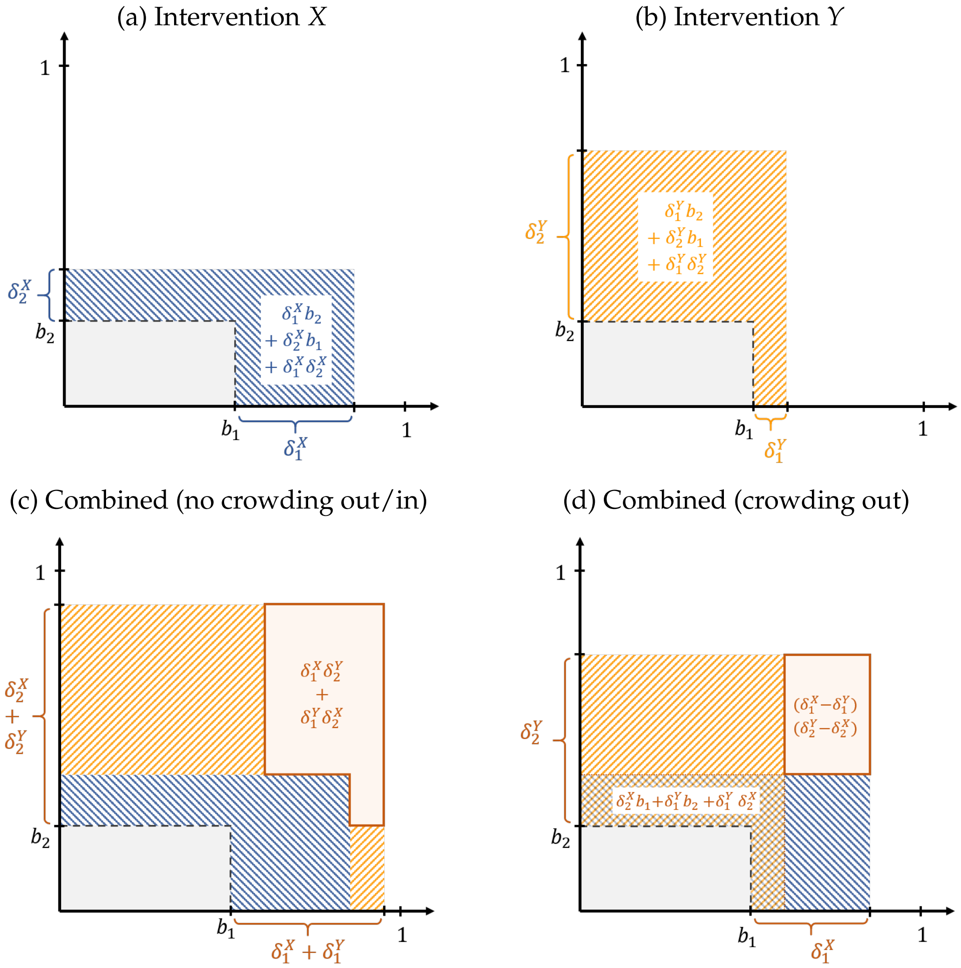
Next, we look at a case where, in each dimension, only the dominant intervention matters, i.e. $\delta_i^{X+Y} = \max(\delta_i^X, \delta_i^Y)$. This is illustrated in Figure 2d. This case is less favorable toward complementarities, as each intervention now only has an impact on one bias dimension, and the condition becomes

$$\Delta B^{XY} - \Delta B^X - \Delta B^Y = (\delta_1^X - \delta_1^Y)(\delta_2^Y - \delta_2^X) - (\delta_2^X b_1 + \delta_1^Y b_2 + \delta_2^X \delta_1^Y) \tag{7}$$

This term is is positiv, if the top right rectangle in Figure 2d, which represents the policy lever complementarity, is larger than the cross-shaded intersection of $X$ and $Y$, which represents loss in impact from $X$ and $Y$ in isolation. Complementarity is more likely the more specialized each intervention is, as the interaction is increasing in $b_1^X$ and $b_2^Y$ and decreasing in $b_1^Y$ and $b_2^X$.

*Complementarity in behavioral outcomes.* — So far we have focused on mechanisms of complementarity in manipulating $B$. How this maps into observable behavior depends on the mapping of $B$ to $e$. The condition for overall policy complementarity in the out-

Figure 2: Depiction of example interventions



(a) Intervention $X$

(b) Intervention $Y$

(c) Combined (no crowding out/in)

(d) Combined (crowding out)

*Notes.* Figures (a) and (b) illustrate the bias mitigation effect of interventions $X$ and $Y$ in isolation, respectively. Figure (c) illustrates their combined effect when their individual effects in each dimension are additive, i.e. there is neither crowding out nor crowding in. Figure (d) illustrated their combined effect when there is perfect crowding out of the less effective intervention in each dimension.

come of interest, energy consumption, can be written as

$$\Delta e^{X+Y} \leq \Delta e^X + \Delta e^Y \tag{8}$$

Typically, one would expect a decreasing responsiveness, as resource consumption is more inelastic at lower levels (e.g. due to a desire for satisfying basic needs like hygiene), so the scope for further conservation effects diminishes with every intervention that is piled upon another. In our framework, this corresponds to function $f^{-1}$ being convex.[11] Intuitively, the more the agent already reduces her consumption, the more difficult it be-

[11] For example, if $V$ has a positive third derivative and the cost function $C$ is linear or quadratic, then $f^{-1}$ is strictly decreasing and convex. A positive third derivative is often labeled prudence and implies a desire for precautionary saving in choice under risk. Of course, $f^{-1}$ could in principle also be concave, so marginal returns are increasing, but this seems implausible. For example, concavity can imply that conservation programs have larger effects for low-baseline consumers, although the opposite is usually true.

comes to further reduce it. Thus, under this assumption, observing complementarities in behavioral outcomes implies complementarities in bias mitigation.

Empirically identifying complementarities is important for optimal policy. Consider the following stylized application: A policy maker has the objective of reducing the average energy consumption $\bar{e}$ in the population and has at her disposal two equally-costly and equally-effective behavioral interventions $X$ and $Y$. Suppose that the budget allows for treating fraction $\alpha \in (0, 1)$ of the population with an intervention. Alternatively, the policy maker can also treat $\alpha/2$ of the population with a combined intervention $X + Y$. The latter is (weakly) superior to the former precisely when the complementarity condition in equation (8) holds. Thus, it is important to study empirically whether two interventions are complements or substitutes.

## 3. Experimental setup

Our field experiment was conducted from early December 2016 to late February/early March 2017 in a sample of students living in dormitory apartments. Each participant was equipped with a smart meter that measured individual energy and water consumption in the shower over the entire study duration. We then evaluated the effect of two different interventions, real-time feedback and shower energy reports, on resource conservation behavior. To test for complementarity, we further implemented a combined intervention in which subjects received both real-time feedback and shower energy reports.

### 3.1. Recruitment of participants

We selected six student dormitory sites in Bonn and Cologne for our sample, and ran the study from early December 2016 to early March 2017. All dormitory residents were students at the University of Bonn, the University of Cologne, or at various smaller universities in the cities. We recruited our subjects from the pool of dorm tenants living in single-person apartments with private bathroom, as this allows us to precisely measure the resource use of each individual. These students have no direct monetary incentives to conserve energy or water, because they pay a flat monthly rent that includes all utility bills. Hence, any observed conservation response would be solely driven by non-monetary motives and unconfounded by income effects.

To participate in the study, residents had to actively agree based on the principle of informed consent. Two additional criteria were levied: subject should not have lengthy absences planned within the intended study period (except during Christmas vacation), and they should own a smartphone compatible with Bluetooth 4.0, which was necessary for implementing the shower energy reports.

The recruiting process started around mid-October 2016. Posters and flyers informed residents of the selected dormitories about the upcoming study, and our local research assistant teams engaged in door-to-door recruiting. Interested students had to complete an online registration survey to provide required information and to give their consent to the collection and analysis of data on their showering behavior. It was explicitly (and truthfully) stated that we would treat any collected data confidentially and not share it with the dormitory administration. As remuneration, each participant received 20 Euros after completing the study, and ten participants were randomly drawn to receive a 300 Euro cash prize. In total, 406 students registered for the study, out of which 361 met our participation criteria.[12] Ten students subsequently dropped out of the study, either because they moved out of their dorm unexpectedly or because we were not able to contact them again. This leaves us with a final sample of 351 participants.

### 3.2. Smart shower meters and smartphone app

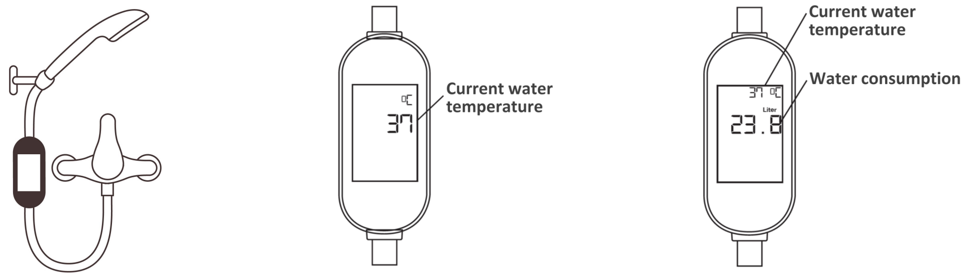
At the beginning of the study, starting on 5th Dec 2016, each participant was equipped with an Amphiro b1 smart shower meter that measures and records data of every water extraction in the shower. The device can be easily attached below the shower head and features a smartphone-sized liquid crystal display, which can be programmed to show various types of information (see Figure 3a). The smart meter is small, lightweight, and needs no battery; power is generated through an integrated hydro turbine, without noticeably affecting water flow in the process. One drawback of the lack of battery is that the device is unaware of the absolute time of day: showers can only be recorded in temporal order, but without time stamps. Once the water flow in the shower starts, the smart meter is powered and begins to measure, among others, the amount of water flowing through, water temperature, and the time passed since beginning of water flow. After water flow is stopped, the device remains powered on for three minutes with the display remaining active. If the water is turned on again within this time frame, the device will continue measurement from the point where it had previously stopped. This accounts for short breaks in water flow when applying soap or shampoo. Once water flow stops for more than three minutes, the device terminates measurement and stores the recorded data as the most recent observation point.

We programmed the shower meters to display select pieces of information to participants in real-time, i.e., while they are taking their showers, contingent on the study progress and assigned experimental condition (as described below). In addition, we asked all participants to install the Amphiro smartphone app around week 5 of the experiment, shortly after the end of the Christmas break. The participants could use the app to upload data from their shower meters via Bluetooth.[13] We were then able to access

[12] The total number of all single apartments in the selected dorms was 1380 (vacancies included), thus our gross recruitment rate was about 30%. For more than half of these apartments, we never encountered the resident, so out of the students we actually managed to talk to, the majority registered for the study.

[13] The process was quite simple. After installing the smartphone app, subjects created an account and paired it to their shower meter. After successful pairing, the meter automatically transmitted all stored data to the app via Bluetooth whenever it was powered on and the smartphone within range.

Figure 3: Amphiro b1 smart shower meter



(a) Position in the shower (b) Display in control mode (c) Display in feedback mode

the uploaded data and use it to create personalized shower energy reports. The original Amphiro smartphone app also calculates summary statistics about users' resource use in the shower, but we deactivated this feature for our study participants, so its only functionality was data uploading. One ancillary benefit of the app was that it stored time and date of each data upload, which allows us to construct approximate time windows for each shower. About three out of four participants (72%) uploaded all data successfully, while the remaining experienced some technical problems. The most common sources of failure were problems with the Bluetooth connection or unexpected incompatibility between smartphone and app. We will come to back to this issue again later.

### 3.3. Implementation of real-time feedback

The live tracking of water use on the shower meter display in feedback mode is what we refer to as real-time feedback, our first type of intervention. We programmed half of the smart meters as control devices and the other half as treatment devices. Control devices only displayed the current water temperature throughout the entire study (Figure 3b). Treatment devices also started in control mode for the first ten showers, which we use to measure baseline behavior, but switched permanently to feedback mode starting from the eleventh shower. In feedback mode, the display shows both the water temperature and the amount of water used (in liters) at any time of the shower (Figure 3c).

### 3.4. Implementation of shower energy reports

Our second type of intervention consists of two personalized shower energy reports. These reports were sent via e-mail and showed descriptive statistics about the subject's water and energy use in the shower, as well as information about environmental impacts. Temperature information was not included, as all subjects received this through their smart meter anyway. To allow for learning about outcomes of single showers, a graphical representation of the subject's history of water use per shower was included. The reports were constructed based on data that was uploaded by subjects through the smartphone app. We sent out additional reminders to upload data before each planned delivery, but the reports themselves were not explicitly announced. Subjects who did not manage to upload any data received a report template with blanks in place of statistical figures and graphs.

Appendix Figure A1 shows the screenshot of a typical shower energy report. After a short introductory text, subjects see a scatter plot of their history of water use per shower since the beginning of the study, including a fitted regression line to help recognize trends and averages. Below the graph, average water use (in liters) and energy use (in kWh) per shower are stated numerically. Furthermore, there is a paragraph with information on projected $CO_2$ emissions per year and the number of trees required to absorb the corresponding amount of $CO_2$. The whole report is formulated concisely in neutral language, to avoid any normative or moral suasion elements. In the second report, we added a social comparison component in the spirit of the original Opower home energy reports, see Appendix Figure A2. Specifically, we assigned a random anonymous peer to each subject and displayed statistics on the peer's energy and water use.[14] At the bottom of each report, there was a personalized link to a mini-survey that we can use to verify if, and how closely, the email has been read.

### 3.5. Experimental design

We implemented a complete 2×2 design with four experimental conditions. Subjects in the control (CON) group received no intervention at all; subjects in the RTF group only received real-time feedback through the smart shower meters; subjects in the SER group only received shower energy reports; and subjects in the DUAL group received both real-time feedback and shower energy reports. Treatment assignment was randomized and the group sizes are as follows: 82 in CON, 88 in SER, 90 in RTF, 91 in DUAL.[15]

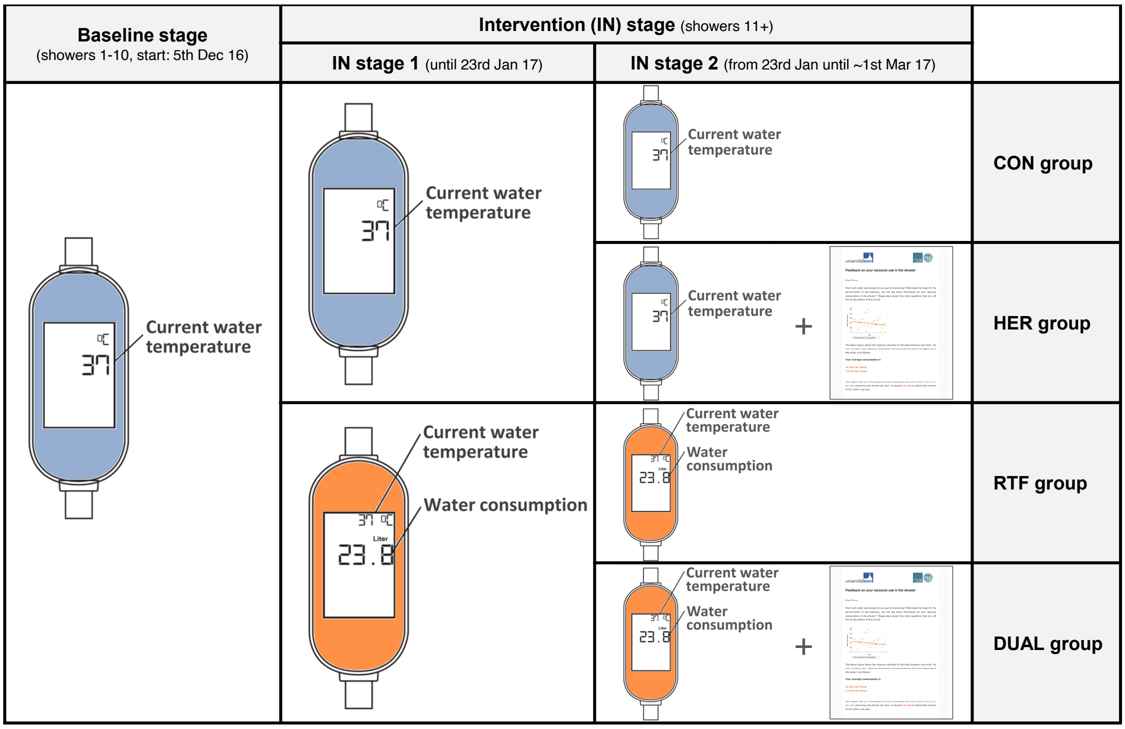
Figure 4 illustrates the experimental design in detail. Each shower meter went through a baseline stage of ten showers, in which it only displayed the current water temperature, regardless of the experimental condition. We use these showers to measure baseline consumption behavior. Starting from the eleventh shower (intervention stage), devices in RTF and DUAL additionally displayed water use in real-time, whereas devices in CON and SER permanently stayed in control mode. About halfway into the study, we started sending energy reports to each subject in the SER or DUAL group; the first report was sent on 23 January 2017 and the second report on 8 February 2017, about two weeks later. We distinguish between intervention (IN) stage 1, in which real-time feedback is switched on but there were no reports yet, and intervention (IN) stage 2, which is the period that begins after the first report was sent out.[16] In order to hold interaction with

[14] The matching procedure was one-sided and ensured that each subject (except the most and the least efficient) was equally likely to see a peer with lower or higher energy use per shower.

[15] For the exact randomization protocol, see Appendix B.

[16] In practice, the distinction between IN stage 1 and 2 is not perfect, as we observe 23 subjects in our sample who had yet to complete all 10 baseline showers when the first report was sent out. If anything, this generates measurement error in our treatment indicators and thus biases estimates toward zero.

Figure 4: Experimental design and timing of interventions


Baseline stage
(showers 1-10, start: 5th Dec 16)
Intervention (IN) stage (showers 11+)
IN stage 1 (until 23rd Jan 17)
IN stage 2 (from 23rd Jan until ~1st Mar 17)
Current water temperature
37
Water consumption
23.8
Liter
+
CON group
HER group
RTF group
DUAL group

experimenters constant, subjects in CON and RTF groups received placebo emails at the exact same time the shower energy reports were sent out. These subjects were simply asked to fill out a mini-survey, the same that came along with the actual reports.

This staggered experimental design allows us to exploit both between- and within-subject variation to cleanly identify and efficiently estimate treatment effects of interest. The effect of real-time feedback in isolation is identified by the comparison between the RTF and CON groups in the (entire) intervention stage, or alternatively by the comparison between the pooled RTF/DUAL group and the pooled CON/SER group in IN stage 1. The effect of shower energy reports in isolation is identified by the comparison between the SER and CON groups in IN stage 2. The additional effect of shower energy reports, when combined with real-time feedback is identified by the comparison between the DUAL and RTF groups in IN stage 2. Differences between the effects of shower energy reports with and without real-time feedback identify policy interaction effects, i.e. whether the two interventions are substitutes or complements. Note that behavior in the CON group may not reflect the "true" counterfactual, as subjects still receive a smart meter with temperature information and placebo emails instead of shower energy report. We would underestimate the effects of our interventions to the degree that subjects respond to this by itself, but any relative comparison across intervention regimes would remain valid.

### 3.6. Behavioral predictions

In order to derive behavioral predictions for each of our experimental groups, we first briefly discuss the channels through which each of the two interventions is likely to work. Our theoretical framework shows that the effect of each regime depends on the degree to which it succeeds in overcoming the aggregate bias, which may be the product of multiple separate factors. Furthermore, real-time feedback and shower energy reports could be complements if they are relatively specialized and operate largely through different channels.

Real-time feedback visually displays live measurement of water use in the shower. This water volume information can debias individuals' beliefs about the amount of water they use, but there is no additional information on energy use or $CO_2$ emissions due to water heating, so severe knowledge gaps about the environmental relevance of showering may remain. In addition, the steadily upward moving liter count is likely to significantly reduce inattention and self-control problems, as users are constantly facing the smart meter display, and the previously abstract and elusive notion of resource use suddenly becomes salient and palpable, infused with a sense of immediacy. It may also facilitate experimentation with various conservation strategies by keeping track of progress in real-time. As the RTF condition in our experiment is essentially a replication of the intervention by Tiefenbeck et al. (2018), albeit more minimalistic and in a sample without monetary incentives, we also expect to find comparable conservation effects.

**Prediction 1.** *Providing real-time feedback through the smart shower meter display in treatment RTF leads to a reduction in water and energy consumption in the shower.*

Shower energy reports provided personalized information about subjects' water use in the shower as well as additional information about energy use and $CO_2$ emissions. We therefore expect that the reports can help close knowledge gaps in these areas and thereby induce conservation behavior, since past evidence suggests that individuals tend to grossly underestimate the energy intensity associated with water heating (Attari et al., 2010). The second report also included a comparison with a randomly assigned and anonymous peer, which might further add motivation through social norms, although Tiefenbeck et al. (2018) find no effect of including comparisons with the co-resident in a two-person household. However, the reports are not immediately salient while showering.

**Prediction 2.** *Providing information through shower energy reports in treatment SER leads to a reduction in water and energy consumption in the shower.*

The conservation effect of knowledge gains through energy reports could be stifled by remaining barriers like limited attention or self-control problems, which can be more suitably targeted by real-time feedback.[17] Vice versa, the effect of real-time feedback may be attenuated if subjects remain unaware of the energy and carbon intensity of warm water use. If the two interventions indeed work largely through these separate behavioral mechanisms, a combined intervention should leverage all mechanisms at the same time. As we argue in the theoretical framework, shower energy reports and real-time feedback could therefore become complements in the sense that one intervention makes the other more effective when implemented jointly.

**Prediction 3.** *Shower energy reports in IN stage 2 lead to a larger (marginal) reduction in water and energy consumption in the shower for subjects who also receive real-time feedback (treatment DUAL) than for subjects who do not receive real-time feedback (treatment SER).*

# 4. Data and descriptive statistics

## 4.1. Measurement data on resource use behavior

For every water extraction in the shower, the smart meters measured, among others, the volume of water used, its average temperature, and the average flow rate (i.e. volume per time unit). The amount of energy used was then calculated based on volume and temperature data, using the standard engineering formula for heat energy.[18] Every subject had a shower meter installed for the whole duration of the study, starting from early December 2016. At the end of the study, in early March 2017, we retrieved the devices and read out the data manually.[19] In this way, we were able to extract an initial data set of 21,469 showers by 327 participants. Unfortunately, no data could be obtained in 24 cases, either because the device was defective or because subjects never used it, or because subjects simply disappeared without a trace (and their shower meters with them).

A number of data cleaning steps are performed before running the empirical analyses. We briefly describe the most important steps here; a more detailed documentation can be found in Appendix C. First, we drop the very first data point of each participant, as they usually started with a test run to check if the device was working. Following Tiefenbeck et al. (2018), we further drop any water extraction with volume below 4.5 liters (in total $2,942$ extractions), as these are unlikely to be actual showers but rather minor extractions for other purposes such as cleaning. As there are rare cases in which the device can pro-

[17] In principle, it is possible that participants also become more attentive about resource use even without visual aid through the smart meter, as would be predicted by rational inattention models when updates in beliefs about environmental impacts are sufficiently large. However, if there is such an effect, it may prove short-lived once reports fade out of memory and resolutions cool off (Allcott and Rogers, 2014; Schwartz and Loewenstein, 2017).

[18] The formula for energy use of water heating is $Q = m \times c_p \times \Delta T$, with heat energy $Q$, mass of water $m$, heat capacity $c_p$ , and $\Delta T$ the difference between the measured water temperature and cold water temperature (assumed to be 12 degrees Celsius). Following Tiefenbeck et al. (2018), we also assume boiler efficiency losses of 35% and distribution losses of 24%.

[19] We already started retrieving some devices in late February, but as the retrieval process was drawn out over a period several days, the end of the study was in early March for most subjects.

duce errors when storing data, we further remove 37 extreme outlier points, defined as such by being more than 4.5 times the subject-specific interquartile range away from the closest quartile.[20] We further exclude 1 device with generally erratic data, 5 devices with fewer than 10 recorded extractions, as well as 3 devices with an abnormally large baseline consumption of 168 liters or more per shower, which is about 40 liters (1.5 standard deviations) away from the rest of the field. In 8 cases, the integrated temperature sensor became defective after some time, and we impute missing information with the average temperature of showers taken while the sensor was still intact. The final data set used for our empirical analyses includes $17,942$ showers by 318 participants.

The shower meter stores the temporal order of showers, so we can easily classify each shower into baseline or intervention stage, as real-time feedback (in the RTF and DUAL groups) started from the eleventh shower. Assigning showers to intervention stage 1 (pre-reports) or stage 2 (post-reports) is slightly more tricky, as the device has no counter for global time. Fortunately, the smartphone app stores the date and time of each data upload, which allows us to construct bounds for when a shower took place. We instructed subjects to use the smartphone app regularly starting from 11 January 2017, and sent additional reminders before each energy report was sent out. Using this timing information, we classify observations into pre-report showers (IN stage 1) or post-report showers (IN stage 2). If there are multiple showers within the range of uncertainty around report dates, we use the switching point implied by constant shower frequency. One complication is that we do not know the timing of showers by the subjects who did not manage to upload any data to the app. Therefore, we impute the timing of showers for these non-uploaders based on the assumption that timing of shower energy reports follows the same distribution for uploaders and non-uploaders. To operationalize this, we use timing information from uploaders to estimate the probability that a shower took place after receiving the first (second) report, and then assign the implied post-report probabilities to showers of non-uploaders. Figure A3 in Appendix A plots the estimated CDFs.[21]

## 4.2. Survey data

To supplement our behavioral data on resource use in the shower, we administered several questionnaires. In the baseline survey, we collected information on individual characteristics (i.e. age, gender, etc.), perceived water use in the shower, shower comfort (i.e. how much they enjoy showering), environmental attitudes and beliefs, as well as a number of personality attributes (i.e. Big Five, patience, etc). In the post-intervention survey, we again collected self-reported data on perceived water use, shower comfort, and environmental attitudes. Furthermore, we administered mini-surveys with each energy

[20] We are particularly strict in only excluding the most implausible data points here. Conventionally, 1.5 or 3 times the interquartile range (IQR) are used as criterion for outliers. For a normal distribution, 4.5 times the IQR away from the nearest quartile corresponds to 6.745 standard deviations away from the mean.

[21] For more details of the imputation procedure, see Appendix D.

Table 1: Descriptive statistics – baseline showers

| | Mean | Std. dev. | 10th pctile | Median | 90th pctile | Obs. |
|---|---|---|---|---|---|---|
| Energy use [kWh] | 2.21 | 1.91 | 0.43 | 1.71 | 4.58 | 2489 |
| Volume [liter] | 37.82 | 30.45 | 9.20 | 29.60 | 76.00 | 2489 |
| Duration [min] | 7.00 | 5.01 | 1.96 | 5.83 | 13.01 | 2489 |
| Temperature [Celsius] | 36.16 | 5.22 | 32.00 | 37.00 | 40.00 | 2463 |
| Flow rate [l/min] | 5.71 | 2.45 | 2.80 | 5.40 | 9.10 | 2489 |

Includes only showers taken in the baseline stage, i.e. first 10 showers and before shower energy reports were sent out. For temperature statistics, devices with broken temperature sensors are excluded. Duration is net of any breaks and calculated by dividing water volume by flow rate.

report, in which subjects were asked to estimate their resource use in the shower.

We mainly make use of information on water use perceptions, shower comfort, and environmental attitudes, and how they change in response to our interventions. Environmental attitude is elicited using four items about pro-environmental behavior and identity, e.g. "I do what is right for the environment, even when it costs more money or takes more time".[22] Shower comfort is elicited using five items on how much subjects enjoy showering, e.g. "I find it relaxing to take a shower".[23] We create indices for shower comfort and environmental attitude, respectively, by taking the simple average of the individual's responses to the relevant items (rated on a 4- or 5-point Likert scale) and then normalizing to mean 0 and standard deviation 1. For perceived water consumption, we asked subjects to estimate how many liters of water they typically use when taking a shower. These estimates can then be directly compared to their actual water use as measured by the smart meter. Note that we refrained from eliciting subjects' beliefs about energy use and carbon emissions from water heating, because we did not want to raise awareness about these issues and risk undermining the shower energy report treatments.

### 4.3. Sample characteristics and baseline behavior

All participants in the field experiment were students at universities in Bonn or Cologne living in single-person dorm apartments, so our sample is rather homogeneous. From the 318 participants represented in our main dataset, 203 lived in a dorm in Bonn and 115 lived in a dorm in Cologne. The share of females was 61 percent.[24] Average age was 23.8 years (median 23 years), with students from all stages of their studies being represented in our sample.

[22] The other items are "Environmental friendliness is part of my personal identity", "How often do you try to conserve water?", and "How often do you try to conserve energy?". We also include a set of questions adapted from Nolan et al. (2008) in the baseline questionnaire.

[23] The other items are "I like showering", "For me, taking a shower is just a means to an end", "I like to let my mind wander when I shower", and "I try to shower as quickly as possible".

[24] In 2016/17, the overall share of female students was 55% at the University of Bonn and 60% at the University of Cologne, suggesting that there was no substantial gender-based selection into our study.

Using the nine showers (the first being excluded) in the baseline stage, where only the current water temperature was displayed, we can construct measures of each subject's baseline resource use behavior. Table 1 presents descriptive statistics about baseline energy and water use per shower, as well as shower duration (net of breaks), water temperature, and flow rate. Shower duration is calculated from dividing water volume by average flow rate. On average, showers in the baseline stage feature 7 minutes of water flow, which amounts to 37.82 liters of water. On average, water is heated up to a temperature of 36.16 degrees Celsius, resulting in energy use of 2.21 kWh per shower. There is substantial variation across showers, as observed from the standard deviations and different quantiles of the distributions. Water and energy consumption follow a right-skewed distribution, thus the median energy use per shower (1.71 kWh) is substantially lower than the mean. The average flow rate of 5.74 liters per minute is relatively low, likely due to dorm infrastructure not being up to modern standards (flow rates of 10-12 liters per minute are more typical for German households).

### 4.4. Randomization checks

Our identification strategy relies on randomization producing treatment groups that are comparable with regard to observable and unobservable subject characteristics. Although it is naturally impossible to test the latter, we can check balance on observable baseline characteristics. Panel A of Table 2 shows results from regressing various measures of subjects' baseline behavior on assigned treatment groups. The differences between groups are very small and treatment assignment is insignificant for predicting any of the behavioral measures, so randomization seems to have worked well. We also check for balance along background characteristics and survey responses (see Table A1 in Appendix A), and again find that treatment assignment is statistically insignificant. Importantly, self-reported environmental attitude and shower comfort are comparable across groups.

### 4.5. Number of showers

On average, we observe 56.8 showers per individual over roughly 12 weeks of our study, which corresponds to a frequency of about two showers every three days. However, the net frequency (i.e. adjusting for absences) might be closer to one shower per day, as our study period included a two weeks Christmas break. In Panel B of Table 2, we check whether the number of showers per individual differs across experimental conditions, but we find that treatments have no effect on the number of showers ($p = 0.669$). Hence, our interventions do not seem to induce adjustments along the extensive margin, and we do not need to worry about subjects compensating shorter showers with more showers, substituting behavior to other facilities (e.g. wash basin, gym showers), or about them compromising on basic hygiene needs. This means that we can make use of the full panel

Table 2: Randomization checks and extensive margin responses

| | Panel A. Baseline averages by individual | | | | | Panel B. |
|---|---|---|---|---|---|---|
| | Energy use<br>[kWh] | Volume<br>[liter] | Duration<br>[min] | Temperature<br>[Celsius] | Flow rate<br>[l/min] | Number<br>of showers |
| SER group | -0.066 | -1.901 | 0.181 | 0.959 | -0.435 | 3.393 |
| | (0.220) | (3.468) | (0.548) | (0.608) | (0.320) | (5.226) |
| RTF group | -0.111 | -1.253 | 0.284 | 0.086 | -0.124 | -2.312 |
| | (0.215) | (3.427) | (0.597) | (0.595) | (0.370) | (5.183) |
| DUAL group | -0.057 | -0.910 | 0.213 | 0.320 | -0.165 | 3.224 |
| | (0.226) | (3.575) | (0.581) | (0.560) | (0.358) | (5.861) |
| Constant | 2.237 | 38.316 | 6.797 | 35.681 | 5.832 | 55.312 |
| | (0.163) | (2.539) | (0.411) | (0.447) | (0.240) | (3.698) |
| Observations | 316 | 316 | 316 | 314 | 316 | 318 |
| R-squared | 0.001 | 0.001 | 0.001 | 0.011 | 0.005 | 0.005 |
| F-test: $p$-value | 0.966 | 0.958 | 0.969 | 0.356 | 0.571 | 0.669 |

Robust standard errors in parentheses. The omitted category is the CON group. For two participants, the device was not able to record information on baseline showers, but we could extract valid data on showers in later stages; hence the number of observations is only 316 in most columns. In addition, two participants with initially defective temperature sensors are excluded in column 4.

structure of our data and analyze (intensive-margin) water and energy conservation effects at the level of individual shower observations.

### 4.6. Presence of imperfect information and behavioral biases

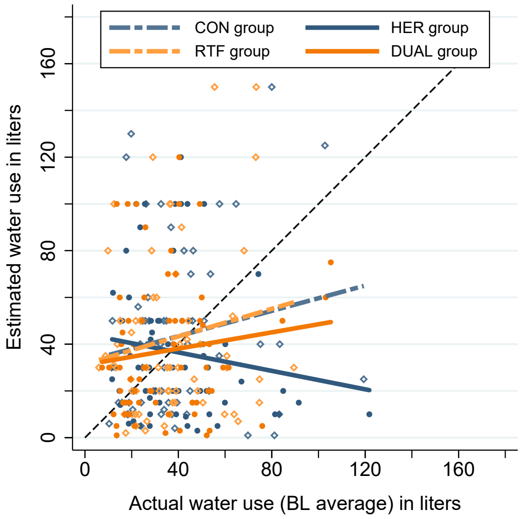
Before moving on to the analysis of our experimental interventions, we provide some descriptive evidence that individuals' resource consumption in our setting may indeed be subject to significant behavioral frictions due to imperfect information and limited attention.

First, we make use of the pre-intervention questionnaire and compare subject's perceptions of their own water use per shower to their actual baseline water use as measured by the smart meter. Figure 5 shows that subjects' estimates are all over the place, and we cannot reject the null hypothesis that estimated and measured water use are in fact uncorrelated (Pearson's $\rho = 0.08, p = 0.1825$). This clearly demonstrates that subjects were not well informed about their own behavioral outcomes prior to any intervention.[25] Interestingly, however, the mean estimate across all subjects (39.8 liters) is close to the actual mean water use per shower in the baseline stage (37.8 liters). This is reminiscent of a "wisdom of crowds" phenomenon and suggests that, on average, our interventions should not work through debiasing beliefs about water use.

Furthermore, subjects are probably especially unaware of how much energy is con-

[25] We excluded 35 subjects who responded to the baseline survey more than 2 weeks after we distributed shower meters, as they have likely reached the intervention stage by then. We also exclude 3 extreme outliers with estimates above 200 liters. The corresponding regression results are presented in Appendix A Table A5.

Figure 5: Pre-intervention awareness about water use per shower



*Notes.* This figure compares estimated water use from the baseline survey with actual water use in the baseline stage (showers 2 to 10), excluding late survey responders. 3 outliers with estimates between 200 and 600 liters are excluded. Point clouds consist of individual observations (hollow diamonds for CON and RTF, solid circles for SER and DUAL) and lines represent separate regression fits for each treatment group. The dashed line starting from the origin is the 45 degree line.

sumed, and hence $CO_2$ emitted, in a typical shower. Attari et al. (2010) show that consumers are in general highly prone to underestimating the amount of energy required for heating up water (e.g., water boilers, dishwashers). We did not elicit beliefs about energy intensity or carbon emissions in the original experimental sample, to avoid the risk of undermining our shower energy report treatments. We did, however, elicit beliefs about carbon emissions in a different sample of students living in the same dormitories three years after the original study ($n = 329$). Without additional information, these students underestimated the carbon impact of warm water use in the shower by a factor of 8 to 9 on average, even though the average guess for the amount of water used per shower was fairly unbiased.[26] Thus, there might be a large potential for encouraging energy conservation through the information provided in shower energy reports.

Although anecdotally compelling, finding direct evidence for inattention or self-control problems in the shower is more tricky. The closest we have is a baseline survey item on how much subjects agree with the statement "I like to let my mind wander when I shower." on a five-point Likert scale. 59% of our sample states that they agree or strongly agree to the statement (34% agree, 25% strongly agree), whereas only 18% of subjects disagree or strongly disagree (13% disagree, 5% strongly disagree), indicating that a lack of focus while showering is prevalent. We further find that subjects' response to this item is significantly correlated with their baseline energy use in the shower (Pearson's $\rho = 0.17$,

[26] On average, students estimated that a typical shower causes emissions of 91.3 grams of $CO_2$ (median 35 grams). The actual emissions amount based on the data from our experiment is about 800 grams. The average guess for amount of water used per shower was 40.4 liters. The survey was conducted in Nov/Dec 2019 among 329 residents of the exact same student dorms in which the original study took place in 2016/17. Only 4 surveyees had already participated in the original study. For more details, see Appendix E.

$p = 0.003$). In fact, it is the single most predictive item for baseline consumption in the entire survey. Our interventions could thus help reduce energy use by reminding subjects not to lose track of time completely under the shower.

## 5. Estimation approach

Next, we describe our strategy for estimating the effects of our interventions on resource use in the shower. The empirical results will be presented in the following section.

### 5.1. Basic estimation strategy

To formally estimate the effects of different intervention regimes, we exploit the staggered introduction of real-time feedback and shower energy reports in the experimental design, which gives us a double-layered difference-in-differences setup. The differential changes in consumption behavior across conditions from baseline stage to intervention stage 1 identify the causal effect of real-time feedback (RTF/DUAL versus CON/SER), and the additional changes from intervention stage 1 to stage 2 identify the causal effect of shower energy reports, both in isolation (SER versus CON) and in conjunction with real-time feedback (DUAL versus RTF).

For estimating the effect of real-time feedback in isolation, the most straightforward and easy-to-interpret approach is to simply compare subjects in the RTF and CON groups over the entire experimental period, as these subjects never received shower energy reports in any form. We do so by estimating the equation

$$y_{it} = \alpha_i + \beta_0 IN_{it} + \beta_1 IN_{it} \times T_i^R + \varepsilon_{it} \,, \tag{9}$$

where the outcome variable $y_{it}$ is energy use (water use) by individual $i$ for shower number $t$, $\alpha_i$ is the individual fixed effect, $IN_{it}$ is an indicator that takes the value 1 if observation $it$ falls into the intervention stage (i.e. $t > 10$), and $T_i^R$ is an indicator for being assigned to treatment group RTF. The coefficient of interest is $\beta_1$, which corresponds to the average treatment effect of real-time feedback (in isolation) over the entire three months of the study. In this specification, we do not have to deal with issues relating to non-compliance and timing of reports, though it comes at the cost of disregarding half of the sample in intervention stage 1.

To make use of the full sample when estimating the effect of real-time feedback, we can compare differential changes in consumption behavior from baseline stage to intervention stage 1 for the pooled RTF/DUAL group versus the pooled CON/SER group, because real-time feedback had already phased in but shower energy reports had not. For intervention stage 2, when shower energy reports started flying in, we split up the pooled groups again, so the regression equation is

$$\begin{aligned} y_{it} &= \alpha_i + IN_{it} \times \left(\beta_0 + \beta_1 T_i^{R/D}\right) \\ &\quad + IN_{it}^{s2} \times \left(\gamma_0 + \gamma_1 T_i^{R/D} + \gamma_2 T_i^S + \gamma_3 T_i^D\right) + \varepsilon_{it}\,. \end{aligned} \tag{10}$$

$IN_{it}$ is again the indicator for the intervention stage, and $IN_{it}^{s2}$ is an indicator for showers that fall into intervention stage 2 (post-report). $T_i^{R/D}$, $T_i^D$ and $T_i^S$ are treatment group indicators, where superscript $R/D$ denotes the combined groups RTF and DUAL, superscript $D$ denotes the DUAL group, and superscript $S$ denotes the SER group only. As $IN_{it}$ remains switched on for the entire intervention period, $IN_{it}^{s2}$ comes on top of that, so all $\gamma$-parameters need to be interpreted as incremental changes from intervention stage 1 to intervention stage 2.

Equation (10) incidentally also includes estimates for the effect of shower energy reports ($\gamma_2$ and $\gamma_3$), but one concern here is that they do not control for differences between RTF and DUAL or between CON and SER in the first intervention stage. Although the pooled groups in intervention stage 1 should behave the same before reports are sent out, some random differences are likely to exist, and these would propagate to the estimates of $\gamma_2$ and $\gamma_3$. For estimating the effects of shower energy reports we therefore prefer the more flexible model in which treatment groups are considered separately from the beginning of the intervention stage:

$$\begin{aligned} y_{it} &= \alpha_i + IN_{it} \times \left(\beta_0 + \beta_1 T_i^{R/D} + \beta_2 T_i^S + \beta_3 T_i^D\right) \\ &\quad + IN_{it}^{s2} \times \left(\gamma_0 + \gamma_1 T_i^{R/D} + \gamma_2 T_i^S + \gamma_3 T_i^D\right) + \varepsilon_{it}\,. \end{aligned} \tag{11}$$

Given the model formulation, we can interpret $\beta_1$ as treatment effect of real-time feedback on energy (water) use per shower in the first stage of the study, while $\gamma_1$ is the change in treatment effect in the second stage. $\gamma_2$ is the treatment effect of shower energy reports in isolation, and $\gamma_3$ is the additional effect of adding shower energy reports to real-time feedback. The relevant comparisons of interests are between SER and CON on the one hand — for the effect of reports without real-time feedback — and between DUAL and RTF on the other hand — for the marginal effect of adding reports to reinforce the already existing real-time feedback.

### 5.2. Estimating treatment effects on the treated

One complication in estimating the effect of shower energy reports is that 28% of subjects did not succeed in uploading any data to the Amphiro smartphone app before we sent out the reports, mostly due to technical problems (e.g., Bluetooth connection failure).[27] For these "non-uploaders", we were unable to provide informative shower energy reports. As the emails were generated automatically, non-uploaders in SER and DUAL groups received report templates with blanks where it was supposed to show statistics on resource use and environmental impacts. Effectively, this leads to imperfect treatment take-up of shower energy reports, although being less the result of deliberate non-compliance than unfortunate circumstances. For participants in the CON and RTF groups, it is inconsequential whether they successfully uploaded data.

One possible approach to estimate treatment effects under imperfect treatment take-up is to run an intention-to-treat (ITT) analysis, which ignores that some participants did not actually receive informative shower energy reports and simply uses treatment assignment to estimate treatment effects. However, this is not very appealing in our context, as failure of information provision due to technical problems is in principle an avoidable problem. The more policy-relevant treatment effect is the effect of delivering informative shower energy reports. Therefore, our preferred approach is to estimate the treatment effect on the treated (TOT), i.e., on subjects who managed to upload data and thus received actual information through the shower energy reports.

The first way in which we estimate the TOT is by simply comparing only the uploaders in SER and DUAL groups with subjects in the CON and RTF groups. The usual concern at this point would be that treatment take-up is not random. Fortunately, our setting limits potential endogeneity concerns for three reasons. First, we include individual fixed effects, so our estimates would still be unbiased if differences between uploaders and non-uploaders do not interact with the treatment effect. Second, subjects only knew that they should use the smartphone app to upload data, but we did not announce that we would use this data to construct shower energy reports. Thirdly, the main cause for non-compliance is not the lack of willingness to use the smartphone app, but unexpected technical failure, which is unlikely to be selected on the trend. To alleviate the most blatant endogeneity issue, we also exclude non-uploaders in the CON and RTF groups who did not report any technical problems.

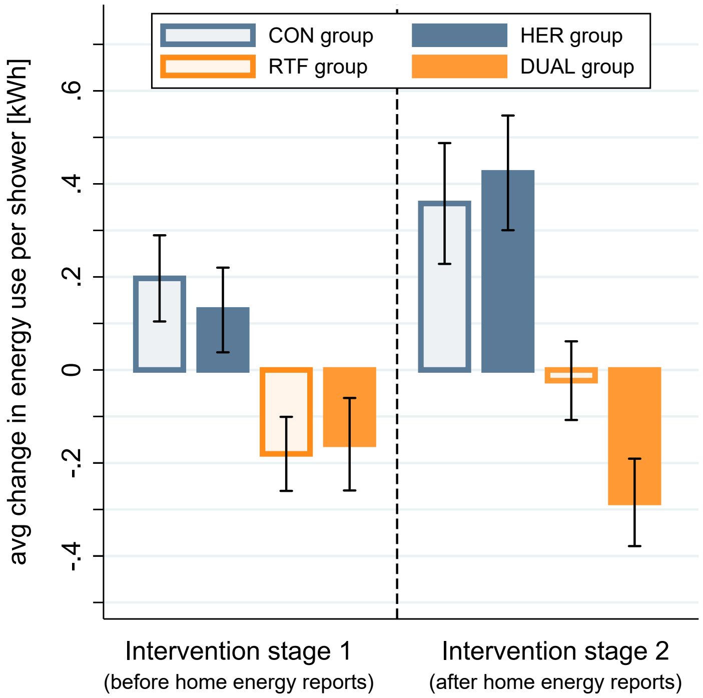
The second way in which we estimate the TOT is by using random treatment assignment as instrument for actual take-up.[28] This can be shown to identify the so-called local average treatment effect (LATE), i.e., the average treatment effect for the sub-population of compliers, in our case the uploaders (Imbens and Angrist, 1994).[29] Compared to the "uploaders-only"-approach, the instrumental variables approach is always consistent,

[27] Out of the 90 non-uploaders in our estimation sample, 63 have explicitly contacted us for technical problems encountered during their upload attempts.

[28] To do this, we create new treatment indicators for the DUAL and SER groups that took the value 1 for showers in IN stage 2 by subjects who were assigned to the respective group *and* who uploaded data through the smartphone app that we could use to construct their shower energy reports. The previously defined ITT indicators are then used as instruments for these new indicators for receiving actual shower energy reports.

[29] This identification result holds under the condition that there are no "defiers", subjects who always do the opposite of what they are prescribed. This monotonicity condition holds by design in our study, because we control the eligibility of shower energy report treatment, so any participant in the sample can be classified either as complier or as never taker in the LATE framework.

Figure 6: Descriptive evidence on energy conservation effects



*Notes.* The bars represent changes in average energy use per shower compared to the baseline period. The error whiskers show standard errors of the mean. Non-uploaders in SER and DUAL as well as non-uploaders without technical problems in CON and RTF are excluded.

but potentially inefficient. We will report the results from both TOT-approaches, but the estimates are very similar, suggesting that endogeneity is not a large issue in our setting.

## 6. Empirical results

### 6.1. Main results

First, we present descriptive evidence on the conservation effects of our interventions. Figure 6 shows subjects' average changes in energy consumption per shower in intervention stage 1 (pre-report) and intervention stage 2 (post-report) compared to the baseline period. The differences-in-differences across treatment groups then correspond to the average treatment effects. In order to show the TOT for shower energy reports, we use the uploaders-only approach of excluding non-compliers in SER and DUAL as well as non-compliers without technical problems in CON and RTF. The graph essentially summarizes our main results in eight bars.

The four bars to the left of the dashed vertical line represent the change in energy use per shower in intervention stage 1 compared to the baseline stage. We can see that relative to subjects in the CON and SER groups, subjects in the RTF and DUAL groups with real-time feedback reduced their energy consumption drastically, by almost 0.4 kWh per shower. Recall that there were no shower energy reports yet at this point.

The four bars to the right of the dashed vertical line represent the change in energy use per shower from baseline stage to intervention stage 2, after shower energy reports were sent out. The first observation is that average energy use in the control group further increased, which could be driven by weather effects, by pending exams leaving students stressed and in need for a long and warm shower, or by Hawthorne effects that decrease over time (Tiefenbeck, 2016).[30] The second observation is that the RTF group and the CON group followed a more or less parallel trend from intervention stage 1 to stage 2, hence the effect of real-time feedback in isolation remains nearly constant at around 0.4 kWh per shower. The third observation is that providing shower energy reports in isolation does not seem to result in effective behavioral change: energy consumption of subjects in the SER group followed the CON group in close synchronization. In light of this, the fourth and final observation is particularly striking: shower energy reports are highly effective when combined with real-time feedback. In fact, subjects in the DUAL group are the only ones to defy the general upward trend and reduce their consumption considerably compared to subjects in the RTF group.

Our descriptive results presented in Figure 6 are confirmed by formal empirical estimates based on the empirical strategy outlined in the previous section. We first focus on estimating the effect of real-time feedback in isolation, before turning to the effect of shower energy reports, for which we need to account for imperfect compliance.

The cleanest way to estimate the effect of real-time feedback is to only compare subjects in the RTF and CON groups over the entire intervention period, by estimating equation (9). Table 3 columns 1 and 2 show that real-time feedback in isolation reduces resource use by 0.40 kWh of energy and 6.3 liters of water per shower compared to the CON group, which corresponds to about 17-18% of baseline use. Columns 3 and 4 present the results from estimating equation (10) on the full sample, using treatment assignment as the independent variable. Subjects in RTF and DUAL conserved about 0.31 kWh of energy and 4.6 liters of water per shower in intervention stage 1, compared to subjects in CON and SER. These are slightly lower than the estimates in columns 1 and 2, partly due to the inclusion of the DUAL and SER groups, partly due to the conservation effect increasing in intervention stage 2, albeit statistically insignficantly.

**Result 1.** *Real-time feedback through the smart meter display led to a reduction in energy (water) consumption by around 0.3-0.4 kWh (4.6-6.3 liters) or 14-18% per shower.*

With the advent of shower energy reports in intervention stage 2, we split the pairs up into the four separate groups again, which incidentally gives us ITT estimates for the effect of shower energy reports; but as discussed earlier, this misses the policy-relevant effect of actually receiving information through shower energy reports. The ITT estimates for the effect of shower energy reports are neither significant for SER nor DUAL, but the

[30] While the baseline phase fell mainly into an unusually warm and dry December, the main intervention months of January and February saw much higher precipitation. Exam periods at the universities began in mid-February.

Table 3: Effect of real-time feedback and ITT estimates

| | *only RTF & CON* | | *Intention to treat* | |
|---|---|---|---|---|
| | (1)<br>Energy<br>[kWh] | (2)<br>Water<br>[liter] | (3)<br>Energy<br>[kWh] | (4)<br>Water<br>[liter] |
| Intervention | 0.283$^{***}$<br>(0.104) | 4.453$^{***}$<br>(1.597) | 0.179$^{***}$<br>(0.067) | 2.915$^{***}$<br>(1.049) |
| Intervention × RTF/DUAL | -0.397$^{***}$<br>(0.125) | -6.346$^{***}$<br>(1.926) | -0.309$^{***}$<br>(0.087) | -4.628$^{***}$<br>(1.387) |
| IN stage 2 | | | 0.187$^{*}$<br>(0.097) | 3.157$^{**}$<br>(1.441) |
| IN stage 2 × RTF/DUAL | | | -0.071<br>(0.118) | -1.745<br>(1.854) |
| IN stage 2 × SER | | | 0.038<br>(0.130) | 0.147<br>(2.006) |
| IN stage 2 × DUAL | | | -0.133<br>(0.093) | -2.302<br>(1.555) |
| Individual fixed effects | *yes* | *yes* | *yes* | *yes* |
| Clusters | 156 | 156 | 318 | 318 |
| Observations | 8446 | 8446 | 17942 | 17942 |
| $R^2$ | 0.379 | 0.375 | 0.403 | 0.404 |

Columns (1) and (2) only include individuals in the RTF or CON group. Standard errors in parentheses are clustered at the individual level. Permutation-based inference for the main coefficients of interest is depicted in Appendix Figure A4.
$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

point estimates for the DUAL group look quantitatively relevant.

Therefore, we move on to the TOT analyses described in Section 5 to estimate the effect of actually receiving information through shower energy reports on conservation behavior. In Table 4, columns 1 and 2 show the estimates obtained by using the uploader-only approach, in which we estimate regression equation (11) on the restricted sample that excludes non-uploaders in SER and DUAL, as well as non-uploaders in RTF and CON without technical issues. Columns 3 and 4 display the LATE estimates, for which we use random treatment assignment to the SER or DUAL group as instruments for actually uploading data and receiving informative shower energy reports. While the LATE approach is consistent even under strong endogeneity of treatment take-up, the uploaders-only approach is potentially more efficient and still consistent if actual take-up (i.e. uploading data) is as good as random conditional on being willing to upload data.

Both approaches produce nearly identical results, suggesting that endogeneity of treatment take-up is not a major issue. The conservation effect of real-time feedback in isolation is also similar to the ones reported in Table 3. The results show that, contrary to our prediction, shower energy reports had no significant conservation effect in the SER group, and the point estimates even run in the opposite direction. While the null effect is not very tightly estimated, we can rule out energy use reductions of more than 7.5%

Table 4: Treatment on the treated (TOT) estimates

| | *Uploaders-only* | | *LATE* | |
|---|---|---|---|---|
| | (1)<br>Energy<br>[kWh] | (2)<br>Water<br>[liter] | (3)<br>Energy<br>[kWh] | (4)<br>Water<br>[liter] |
| Intervention | 0.179<br>(0.111) | 2.628<br>(1.702) | 0.172$^{*}$<br>(0.102) | 2.533<br>(1.565) |
| Intervention × RTF/DUAL | -0.388$^{***}$<br>(0.134) | -5.753$^{***}$<br>(2.124) | -0.365$^{***}$<br>(0.125) | -5.481$^{***}$<br>(1.981) |
| Intervention × SER | 0.027<br>(0.154) | 0.837<br>(2.415) | 0.016<br>(0.134) | 0.733<br>(2.082) |
| Intervention × DUAL | 0.035<br>(0.113) | 0.576<br>(1.860) | 0.109<br>(0.107) | 2.159<br>(1.751) |
| IN stage 2 | 0.150<br>(0.093) | 2.770$^{*}$<br>(1.422) | 0.189$^{*}$<br>(0.098) | 3.273$^{**}$<br>(1.460) |
| IN stage 2 × RTF/DUAL | -0.021<br>(0.118) | -1.142<br>(1.913) | -0.053<br>(0.120) | -1.463<br>(1.908) |
| IN stage 2 × SER | 0.090<br>(0.137) | 0.714<br>(2.168) | 0.042<br>(0.162) | -0.084<br>(2.510) |
| IN stage 2 × DUAL | -0.222$^{**}$<br>(0.100) | -3.702$^{**}$<br>(1.756) | -0.215$^{*}$<br>(0.116) | -3.836$^{*}$<br>(2.037) |
| Individual fixed effects | *yes* | *yes* | *yes* | *yes* |
| Clusters | 261 | 261 | 318 | 318 |
| Observations | 14712 | 14712 | 17942 | 17942 |
| $R^2$ | 0.413 | 0.415 | 0.004 | 0.004 |

In columns (1) and (2), we exclude all non-uploaders in SER and DUAL as well as all non-uploaders in RTF and CON who did not report a technical problem. In columns (3) and (4), we use treatment assignment to SER and DUAL, respectively, interacted with the IN stage 2 indicator as instrument for receiving informative shower energy reports. The reported $R^2$ in Columns (3) and (4) is the within $R^2$. Standard errors in parentheses are clustered at the individual level. Permutation-based inference for the main coefficients of interest is depicted in Figure A6.
$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

per shower with 90% confidence in the (less precise) LATE specification. Furthermore, we can reject the hypothesis that shower energy reports in isolation were as effective as real-time feedback in isolation ($p < 0.003$ in all specifications).

**Result 2.** *Shower energy reports in isolation did not induce any significant reduction in energy and water consumption per shower.*

Statistical imprecision aside, this does not imply that the shower energy reports are generally ineffective in our setting, but only when administered in isolation to the SER group. In stark contrast, we find that subjects in the DUAL group further reduced energy use by around 0.22 kWh (water use by around 3.8 liters) per shower in intervention stage 2, which corresponds to another 10 percentage points reduction from baseline consumption. This means that adding shower energy reports boosted the effectiveness of real-time feedback by more than 50%. The difference between energy conservation effects in the DUAL group and the SER group is weakly significant in the uploaders-only specification ($p = 0.067$). Unfortunately, we do not have enough power to detect this differential effect with larger statistical certainty, due to the technical issues with the smartphone app. Our results are, however, fully robust to randomization-based inference methods (Young, 2019), as presented in Appendix Figures A4 and A6.

**Result 3.** *Combining real-time feedback with shower energy reports further reduced energy (water) use by around* 0.22 *kWh (3.8 liters) per shower and thus boosted the conservation effect of real-time feedback in isolation by more than* 50%*.*

This contrast between the effect of shower energy reports with and without real-time feedback is all the more remarkable given that subjects in DUAL had already cut their energy consumption per shower significantly in response to real-time feedback and thus had less room for further behavioral adjustments, which is exactly one of the opposing effects against complementarity we described in the theoretical framework. Overall, there seems to be a strong complementarity between real-time feedback and shower energy reports. This is consistent with our theoretical framework, which shows that in the presence of multiple sources of bias to resource conservation, behavioral interventions may need to overcome all significant sources of bias simultaneously in order to unfold their full effect. While shower energy reports provide information about resource use and associated environmental impacts, the lack of salience in resource consumption is likely to hinder conservation efforts. Real-time feedback through smart meters could thus turn environmental considerations into action by putting them into focus while showering. We will analyze the underlying mechanisms more closely in Section 7.

## 6.2. Treatment effect dynamics

We now investigate whether the conservation effects of real-time feedback and shower energy reports remain stable over the three-month period of our study. The previous subsection already documents that the effect of real-time feedback does not drop from the first to the second intervention stage. Therefore, we now focus on the 5-6 weeks period of IN stage 2. To estimate dynamic effects, we extend the empirical model for average treatment effects i.e. equation (11), by interacting with a time variable $Z_i$:

$$\begin{aligned} y_{it} = \alpha_i &+ IN_{it} \times \left(\beta_0 + \beta_1 T_i^{R/D} + \beta_2 T_i^H + \beta_3 T_i^D\right) \\ &+ IN_{it}^{s2} \times \left(\gamma_0 + \gamma_1 T_i^{R/D} + \gamma_2 T_i^H + \gamma_3 T_i^D\right) \\ &+ IN_{it}^{s2} \times Z_i \times \left(\delta_0 + \delta_1 T_i^{R/D} + \delta_2 T_i^H + \delta_3 T_i^D\right) + \varepsilon_{it}\,. \end{aligned} \tag{12}$$

We explore two variants of $Z_i$. In the first variant, we look additionally at energy use per shower after the second shower energy report was sent about two weeks after the first re-

Table 5: Treatment effect dynamics

| | $Z_i = \mathbb{I}\{post\ 2nd\ report\}$ | | $Z_i = \#\ weeks\ after\ 1st\ report$ | |
|---|---|---|---|---|
| | (1)<br>Uploaders | (2)<br>LATE | (3)<br>Uploaders | (4)<br>LATE |
| IN stage 2 | 0.139<br>(0.103) | 0.176<br>(0.110) | 0.065<br>(0.124) | 0.109<br>(0.127) |
| IN stage 2 × RTF/DUAL | -0.027<br>(0.128) | -0.053<br>(0.134) | 0.047<br>(0.156) | 0.019<br>(0.159) |
| IN stage 2 × SER | 0.092<br>(0.148) | 0.048<br>(0.169) | 0.198<br>(0.181) | 0.174<br>(0.202) |
| IN stage 2 × DUAL | -0.068<br>(0.123) | -0.041<br>(0.135) | 0.030<br>(0.166) | 0.075<br>(0.177) |
| IN stage 2 $\times Z_i$ | 0.019<br>(0.093) | 0.022<br>(0.090) | 0.032<br>(0.027) | 0.029<br>(0.026) |
| IN stage 2 × RTF/DUAL $\times Z_i$ | 0.012<br>(0.123) | 0.000<br>(0.119) | -0.026<br>(0.037) | -0.026<br>(0.035) |
| IN stage 2 × SER $\times Z_i$ | -0.002<br>(0.126) | -0.010<br>(0.136) | -0.041<br>(0.042) | -0.051<br>(0.047) |
| IN stage 2 × DUAL $\times Z_i$ | -0.279<br>(0.209) | -0.316<br>(0.215) | -0.099<br>(0.064) | -0.114$^{*}$<br>(0.067) |
| Individual fixed effects | *yes* | *yes* | *yes* | *yes* |
| Clusters | 261 | 318 | 261 | 318 |
| Observations | 14712 | 17942 | 14712 | 17942 |
| $R^2$ | 0.413 | 0.005 | 0.413 | 0.005 |

The results are obtained by estimating equation (12). The full table with all the coefficients is presented in Appendix A Table A3. In columns (1) and (3), we exclude all non-uploaders in SER and DUAL, as well as all non-uploaders in RTF and CON who did not report a technical problem. In columns (2) and (4), we use treatment assignment to SER and DUAL, respectively, interacted with the IN stage 2 indicator as instrument for receiving informative shower energy reports. The reported $R^2$ in Columns (2) and (4) is the within $R^2$. Standard errors in parentheses are clustered at the individual level.
$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

port. In the second variant, we interact each treatment group indicator with a linear time trend, so the $\delta$ coefficients can be interpreted as weekly depreciation (or appreciation) rate of energy conservation effects by intervention regime.

Table 5 shows that the effect of shower energy reports in the DUAL group seems to gradually unfold over time. In fact, the reduction in energy use is not yet statistically significant in the first two weeks of intervention stage 2; columns (1) and (2) show that the average conservation effect is driven largely by the final 3-4 weeks of the study, i.e. after the second reports were sent out. However, this does not seem stem from a discrete jump, but rather from a continuous trend. In columns (3) and (4), we estimate that the conservation effect per shower in the DUAL group *increases* by a rate of around 0.1 kWh every week. We should note these changes over time are mostly statistically insignificant and therefore to be interpreted with caution. Shower energy reports in isolation (SER group) show no signs of any dynamic pattern; the coefficient is identical before and after the second report. The effect of real-time feedback in isolation also appears to stay constant in intervention stage 2, overall showing no signs of weakening within the 3 months of our experiment.[31]

There are several potential explanations for the pattern of increasing behavioral responses over time that we observe in the DUAL group. For one, subjects may have skimmed through the email reports initially and only looked at it more carefully later. What speaks against this explanation is that most of the subjects responded to the attached mini-surveys within few days after we sent out the email and that the overall response rate was much higher in the first than in the second report.[32] Nevertheless, it may well be possible that the apparent increase over time is at least partly due to lower measurement error of when subjects where actually treated. Also, the social comparison in the second report might have provided additional motivation, which then interacted with real-time feedback, as would be predicted by the theoretical framework. A third explanation is that subjects may have required some time to try out and discover new strategies for further reducing energy use. This experimentation channel seems consistent with the finding that subjects in the DUAL group do not conserve energy by reducing their shower duration in the second intervention stage, but rather through adjusting flow rate and water temperature. Importantly, the results speak against pure Hawthorne effects or short-lived attention boosts, as these would rather predict an "action-and-backsliding" pattern (Allcott and Rogers, 2014; Schwartz and Loewenstein, 2017).

### 6.3. Heterogeneous treatment effects

Particular subgroups of individuals may have responded more strongly to our interventions than others. Previous studies often find that households or individuals with high baseline consumption tend to respond more strongly to policy interventions targeted at their conservation behavior (e.g. Allcott 2011; Ferraro and Price 2013; Tiefenbeck et al. 2018). For example, Allcott (2011) reports that Opower home energy reports achieved virtually no savings for households in the bottom decile of baseline energy use, whereas the treatment effect for top-decile users was 6.3% savings. Tiefenbeck et al. (2018) estimate that real-time feedback has an additional conservation effect of 0.31 kWh for a 1 kWh increase in baseline energy use per shower. Policy makers concerned about cost-effectiveness can therefore purposefully target high-baseline users.

To estimate heterogeneity along the dimension of baseline energy use, we extend the

[31] This is consistent with other studies using the Amphiros smart meter. For example, Agarwal et al. (2020) find stable effects for an intervention duration of up to 6 months, as well as evidence for strong persistence several months after the intervention.

[32] In fact, 53% (40%) of all subjects in DUAL responded within one day of receiving the first (second) report, and 80% (48%) did so within one week. Overall response rate was 81% (48%).

Table 6: Treatment effect heterogeneity

| | (1)<br>continuous | (2)<br>$\mathbb{I}\{> median\}$ |
|---|---|---|
| ... | ... | ... |
| Intervention × RTF/DUAL | -0.403$^{***}$ | -0.254$^{***}$ |
| | (0.127) | (0.096) |
| IN stage 2 × RTF/DUAL | -0.014 | 0.171$^{*}$ |
| | (0.117) | (0.102) |
| IN stage 2 × SER | 0.095 | 0.267$^{**}$ |
| | (0.139) | (0.121) |
| IN stage 2 × DUAL | -0.239$^{**}$ | -0.156$^{*}$ |
| | (0.102) | (0.093) |
| ... | ... | ... |
| Intervention × RTF/DUAL × Baseline energy use | -0.164 | -0.247 |
| | (0.119) | (0.266) |
| IN stage 2 × RTF/DUAL × Baseline energy use | -0.094 | -0.385$^{*}$ |
| | (0.101) | (0.228) |
| IN stage 2 × SER × Baseline energy use | -0.021 | -0.368 |
| | (0.124) | (0.268) |
| IN stage 2 × DUAL × Baseline energy use | -0.097 | -0.166 |
| | (0.092) | (0.203) |
| Other treatment variables | *yes* | *yes* |
| Individual fixed effects | *yes* | *yes* |
| Observations | 14675 | 14675 |
| $R^2$ | 0.413 | 0.413 |

The coefficients are obtained by estimating equation (13). For visual ease, not all coefficient estimates are presented. The full table with is can be found in Appendix A Table A4. All non-uploaders in SER and DUAL as well as all non-uploaders in RTF and CON who did not report a technical problem are excluded. Baseline energy use is demeaned, so main effects represent TEs at the sample mean. Standard errors in parentheses are clustered at the individual level. $^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

basic statistical model in equation (11) with interaction terms:

$$\begin{aligned} y_{it} = \alpha_i &+ IN_{it} \times \left(\beta_0 + \beta_1 T_i^{R/D} + \beta_2 T_i^H + \beta_3 T_i^D\right) \\ &+ IN_{it} \times X_i \times \left(\lambda_0 + \lambda_1 T_i^{R/D} + \lambda_2 T_i^H + \lambda_3 T_i^D\right) \\ &+ IN_{it}^{s2} \times \left(\gamma_0 + \gamma_1 T_i^{R/D} + \gamma_2 T_i^H + \gamma_3 T_i^D\right) \\ &+ IN_{it}^{s2} \times X_i \times \left(\mu_0 + \mu_1 T_i^{R/D} + \mu_2 T_i^H + \mu_3 T_i^D\right) + \varepsilon_{it} \end{aligned} \tag{13}$$

where variable $X_i$ is a measure of subjects' baseline energy consumption per shower. As a measure of baseline consumption, we use a subject's average energy use in the 9 baseline showers (the first shower is excluded), re-centered around the sample mean (2.21 kWh) so that intercept terms can be interpreted as effects at the mean. In addition, we report a specification where $X_i$ is an above-median indicator.

Table 6 presents TOT estimates of heterogeneous effects along baseline energy use.

Note that we only show the main coefficients of interests here to keep the table visually tractable, but the full set of coefficients can be found in Table A4 in the Appendix. Consistent with previous literature, we find that the effect of real-time feedback in isolation increases with baseline use. In intervention stage 2, compounding the effects over both periods ($\hat{\lambda}_1 + \hat{\mu}_1$), subjects with 1 kWh higher baseline reduce their energy use per shower by an additional 0.26 kWh ($p = 0.069$) on average. Above-median baseline users (mean 3.30 kWh) save 0.63 kWh ($p = 0.039$) of energy more per shower compared to subjects with below-median baseline use (mean 1.17 kWh). This is consistent with the notion that real-time feedback reduces "slack" in resource use, but does not lead subjects to compromise on basic needs. It also appears that providing information through shower energy reports in the DUAL condition induces about double the conservation effect for above-median users ($\hat{\gamma}_3 + \hat{\mu}_3 = -0.322$ kWh, $p = 0.075$), compared to below-median baseline users ($\hat{\gamma}_3 = -0.156$ kWh, $p = 0.096$) in intervention stage 2, although the difference is not significant ($p = 0.414$). Shower energy reports in isolation (SER group) are neither effective for low- nor high-baseline users. In fact, it seems that subjects with below-median baseline use tend to increase their energy use in intervention stage 2 ($p = 0.028$).

# 7. Underlying mechanisms

The empirical results show that, in our setting, shower energy reports seem to be ineffective in isolation, but induce large and significant conservation effects when combined with real-time feedback, which suggests that our interventions are strong complements. Through the lens of the theoretical framework in section 2, the most plausible mechanism for this finding is that the two interventions operated through complementary policy levers. Shower energy reports may have increased knowledge about environmental impacts of warm water use in the shower, but this in itself may not achieve reductions in energy consumption if subjects still face bias due to limited attention or self-control problems. Real-time feedback could help mitigating these problems and thus enable knowledge gains to translate into conservation behavior. If, on the other hand, shower energy reports and real-time feedback both operated through the same policy levers, we would generally not expect complementarities unless there is some type of crowding in effect, e.g. if the combined intervention leads to positive attention or motivation spillovers. In this section, we conduct a number of analyses to explore the mechanisms underlying our main empirical results.

## 7.1. Awareness about resource intensity and environmental impacts

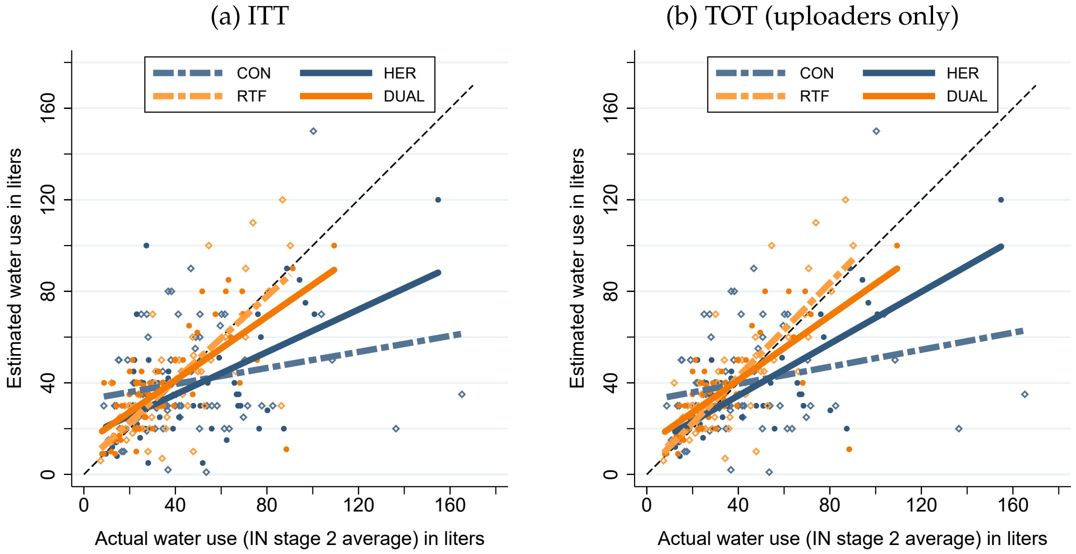
A crucial element of both interventions in our study is that they can enable learning about the outcomes of one's behavior. Real-time feedback through the smart meter provides immediate display of water use (and temperature) for the current shower. Shower energy reports also contain information of individuals' entire history of water (and energy) use per shower since the start of the study, with the difference that it comes in retrospect. Hence, a first manipulation check for our interventions is to analyze their effect on subjects' awareness about their own water use per shower.

In the post-intervention survey at the end of the study, we asked subjects to again estimate the amount of water they typically use per shower. Recall that prior to the interventions, subjects' assessments were virtually uncorrelated with their actual water use, with low-baseline users overestimating and high-baseline users underestimating their water use (see Figure 5). The picture changes completely after the interventions. Figure 7 plots individuals' post-intervention estimates as a function of their average water use per shower as measured by the smart meter. The corresponding regression table A5 is presented in Appendix A. Whereas subjects in the CON group remain as ignorant as before, subjects who received real-time feedback (RTF and DUAL group) are now able to estimate their water use almost without bias, so the fitted regression lines are close to the identity line. While the slope looks slightly flatter for the DUAL group compared to the RTF group, the difference is not statistically significant. Importantly, shower energy reports in isolation (SER group) also induce strong learning effects about water use, as estimated water use increases visibly in actual water use per shower (TOT slope 0.57), and significantly more strongly than in the CON group ($p = 0.025$). We cannot reject that learning through shower energy reports is more effective with real-time feedback than without ($p = 0.497$). While these analyses focus on the bias of subjects' estimates (conditional on actual water use), we obtain similar results when we look at the magnitude of absolute estimation errors across groups. Table A6 in Appendix A shows that subjects in the three treated groups are on average about 27-30 percentage points closer to their actual water use than subjects in the CON group, and notably, the effect is virtually the same for SER, RTF, and DUAL groups.

Taken together, the results show that subjects in our study did engage with the interventions and thereby became more aware of their own water use behavior in the shower. However, belief updates about water use per se are unlikely to drive our main results. First, subjects' prior beliefs about water use were by and large unbiased on average. Second, although the posterior beliefs in the SER group do not become quite as accurate as in the RTF group, we would have expected at least some conservation effect through shower energy reports in isolation if belief updating about water use was the main mechanisms. This points to the importance of the immediacy and salience of the real-time feedback intervention, which can help subjects track their water use while showering and overcome inattention problems.

In contrast to real-time feedback, shower energy reports did not only contain information about water use, but also on energy use and environmental impacts in terms of $CO_2$ emissions. This can explain why subjects in the DUAL group reduced their energy consumption even further after receiving the reports. As a manipulation check for whether

Figure 7: Post-intervention awareness about water use per shower



(a) ITT

(b) TOT (uploaders only)

*Notes.* Both graphs compare subject's water use estimates in the final questionnaire with their actual water use in intervention stage 2. Graph (b) only uses the subsample defined for the uploaders-only approach. 7 outliers with estimates between 200 and 500 liters are excluded. Point clouds consist of individual observations (hollow diamonds for CON and RTF, solid circles for SER and DUAL) and lines represent separate regression fits for each treatment group. The dashed line starting at the origin is the 45 degree line.

subjects responded to this information, we conducted a supplementary survey in a new sample of 329 students at the end of 2019 (see also Section 4.6). After eliciting prior beliefs about water consumption and $CO_2$ emissions per shower, we randomly presented one fact sheet (out of three) to each surveyee, mimicking the basic informational content of our original interventions. The "CON sheet" only reported the average water temperature in the shower, the "RTF sheet" also included the average amount of water used, and the "SER sheet" further added information on energy use and $CO_2$ emissions. After presenting the fact sheets, we elicited posterior beliefs as well as conservation intentions. We find that the SER sheet induces surveyees to drastically adjust their beliefs about $CO_2$ emissions upwards compared to the CON or RTF sheets ($p < 0.001$). This experimentally-induced belief update is further associated with a 0.24 standard deviations ($p = 0.003$) increase in self-stated intention to take shorter showers in the future, compared to the RTF sheet group. For further details, see Appendix E.

Shower energy reports seem to induce knowledge gains about the environmental impact of showering, yet they are only associated with significant conservation effects when combined with real-time feedback. One of the key insights of our theoretical framework is that if multiple sources of bias play a role, different behavioral interventions can become complements, because a single narrowly-targeted intervention is undermined by the presence of other behavioral biases. Hence, our empirical results suggest that, in the absence of real-time feedback, additional barriers like limited attention or self-control problems have prevented knowledge gains and good intentions from translating into actual behavior.

### 7.2. Engagement with shower energy reports

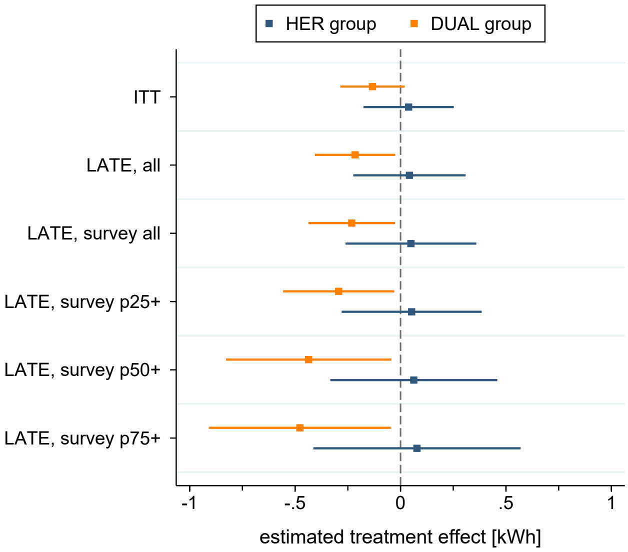
One potential alternative channel is differential treatment engagement, in the sense that subjects in different treatment groups may pay more or less attention to the interventions per se. For example, if previous exposure to real-time feedback induced subjects in the DUAL group to read shower energy reports more carefully than subjects in the SER group, this might lead to complementarity between the two interventions through some type of crowding in or foot-in-the-door effect as described in the theoretical framework. The previous subsection shows that shower energy reports did induce significant learning effects about water use in the shower also in the SER group. Furthermore, we can also directly assess whether the level of scrutiny was similar in the SER and the DUAL group. To do so, we make use of the mini-surveys that were attached to each of the two report emails. As described before, each email included a link to a survey in which we asked subjects to give an estimate of the amount of water they use in a typical shower. The survey link was at the bottom of the email, so subjects had to scroll through all the statistics on resource use and $CO_2$ emissions before clicking on it. We therefore use survey responses as proxy for the level of engagement with the feedback email.

Table A7 in Appendix A shows response rates by treatment group in the uploaders-only sample. Recall that subjects in the RTF and CON groups received Placebo emails containing a link to the same mini-survey. The overall response rates of uploaders was 87% for the first email and 71% for the second email. The share of respondents in the SER group was 8.4%$p$ lower than in the DUAL group for the first email ($p = 0.203$), and 9.4%$p$ higher for the second mail ($p = 0.308$); both differences are statistically insignificant. Apart from the extensive margin, Table A7 further shows subjects' relative estimation error by treatment group, defined as percent deviation of estimated water use in the mini-survey from the actual water use per shower.[33] Smaller estimation errors are an indication of subjects paying closer attention while reading the reports. Respondents in the SER group were only 10% off on average, and they actually gave more precise estimates than respondents in the DUAL group ($p = 0.039$), who were 21% off on average. Notwithstanding, both groups still outperform the CON group (49% off on average) by far. Overall, we find no evidence that uploaders in the DUAL group studied reports more carefully than uploaders in SER group.

As an additional plausibility check that it is not lower level of engagement with the shower energy reports that prevented energy conservation in the SER group, we look at whether subjects who studied the reports more closely also engaged more strongly in conservation actions. For this purpose, we again make use of subjects' water use assessments in the mini-surveys and regress energy use per shower on several new shower

[33] As measure for actual water use per shower, we take the number that was calculated for each subject when sending out the shower energy reports.

Figure 9: Effects for different levels of engagement with shower energy reports



*Notes.* The points represent estimated regression coefficients for the effects of shower energy reports in intervention stage 2, where treatment engagement status is instrumented with treatment assignment (with the exception of ITT). Lines represent 90% confidence intervals. "LATE, survey all" includes all subjects who uploaded data and clicked on the mini survey. The labels "p25+/p25+/p75+" denote the groups of subjects whose estimate precision, defined as distance between estimated and measured water use per shower, was above the 25th, 50th, or 75th percentile of all subjects, respectively.

energy report treatment indicators that increase in their level of strictness. Specifically, we define an indicator for whether subjects uploaded data *and* clicked on the mini survey in their report, and additional indicators for whether a subject's estimate precision, defined as distance between estimated and measured water use per shower, was above the 25th, 50th, or 75th percentile of all subjects, respectively. To avoid the endogeneity issue at hand, we use treatment assignment as instrument for level of engagement with reports. Figure 9 plots the coefficients and confidence intervals for the effect of shower energy reports in SER and DUAL group, respectively. The estimated conservation effect in the DUAL group increases monotonically with the strictness of our compliance definition, reaching almost 0.5 kWh for the strictest indicator. In contrast, even the most studious subjects in the SER group did not reduce their energy use in response to the reports, which corroborates our interpretation that some source of bias such as limited attention may have prevented shower energy reports in isolation from inducing behavioral change.

## 7.3. Other potential mechanisms

There are a number of alternative channels through which our interventions could affect conservation behavior. For one, they could trigger Hawthorne effects, but recall that also subjects in the control group received a smart meter and placebo emails reminding them to upload their data. See Appendix F.1 for a more detailed discussion of why Hawthorne or cueing effects are unlikely to explain our findings. Furthermore, it is possible that real-time feedback and shower energy reports reduced subjects' shower comfort or increased their pro-environmental attitude, but based on survey evidence, neither seems to be the case (see Appendix F.2). If anything, we observe a decrease in self-perceived pro-environmental attitudes in the treated groups compared to the control group, potentially due to feedback provision curbing the capacity for distorted self-image formation.

## 8. Conclusion

In this paper, we argued that if multiple sources of behavioral bias (e.g., imperfect information and limited attention) simultaneously prevent individuals from acting on their values and intentions, then combining interventions that each target a different source of bias can result in complementarity, meaning that each intervention becomes more effective when implemented in conjunction with the other(s) than in isolation. We first introduced a theoretical framework that delineates the interplay of behavioral interventions and illustrates mechanisms for complementarity and substitutability in a setting with multiple behavioral biases; in particular, the potential for complementarity becomes larger the more differentiated the interventions are with regard to their targeted biases. We then presented results from a three-month field experiment on energy conservation behavior in a specific resource-intensive everyday activity (showering), in which we evaluated interaction effects between two types of interventions: shower energy reports, which provided information on energy use and carbon emissions via email, and real-time feedback through a smart meter display, which made water consumption in the shower immediately salient. While only the latter induced a signficant conservation effect when implemented in isolation, combining both interventions resulted in a striking complementarity. It seems that knowledge gains about environmental impacts only triggered conservation behavior once resource use was additionally made salient through real-time feedback.

Although our interventions were targeted towards one specific resource-intensive activity, showering, the effect sizes are also quantitatively meaningful on the aggregate household level, which is all the more remarkable given that our subjects had no monetary incentives to conserve resources. In our study, real-time feedback in isolation lowered consumption by 0.4 kWh (6.3 liters) per shower; adding shower energy reports further lowered consumption by 0.22 kWh (3.8 liters). For comparison, total daily energy use for lighting in German households is about 0.33 kWh per person on average.[34] In his influential evaluation of the Opower home energy reports, which target *aggregate* electricity use in U.S. households, Allcott (2011) finds a household-level conservation effect of 0.62 kWh per day. One limitation of our study is that we do not observe subjects' consumption behavior outside the shower. However, in a related study that uses Am-

[34] Source: German Federal Statistical Office.

phiro smart shower meters in a representative household sample in Singapore, Schmitt et al. (2021) find that the direct conservation effect in the shower may even understate the effect on overall household water consumption. This is in line with recent evidence for potential positive spillover effects of pro-environmental interventions (Jessoe et al., 2021; Sherif, 2021).

We attempted to make a step towards understanding why different interventions can be complements (or substitutes). While both our theoretical framework and our field experiment are tailored to a very specific setting, the notion that potentially multiple different barriers need to be overcome for behavioral change can be relevant in other contexts as well, including situations involving more standard economic barriers such as lack of incentives or constraints on time, money, or technology. Such complexity of behavioral mechanisms is a pervasive feature of many domains of our lives, and it is likely that this creates numerous opportunities for complementarities between different interventions, yet many of these may still be untapped.[35]

Further research is necessary to investigate whether this channel for complementarity of interventions we propose also generalizes to other settings and more representative samples. Nevertheless, our study underlines that any evaluation is inevitably confined to the particular policy and choice environment that consumers act in, which may itself be malleable. Interventions that may seem feeble at first glance may thus be able to unfold their full potential once combined with other interventions that address remaining sources of behavioral bias. For example, our results suggest a special role for interventions that increase the salience of one's resource use: giving individuals simple tools that allow them to track their use may also make their behavior more sensitive to other policies, such as price incentives (Jessoe and Rapson, 2014). Hence, behavioral policy design should not only consider through which channels a particular intervention affects behavior, but also attempt to identify and overcome behavioral barriers that may still remain.

New policies are always introduced to an existing set of policies, institutions, and norms. As social scientists are beginning to pioneer the process from small-scale proof-of-concept studies to large-scale interventions (Banerjee et al., 2017), future research should therefore synchronously advance our knowledge on the interplay of different policy instruments.

[35] Indeed, some empirical findings in the literature are at least suggestive of mechanisms at work that are similar to the one we suggest. For example, Cortes et al. (2019) find that text-message based curricula supporting good parenting practices work less well when parents face high cognitive load than during time periods when the load is lighter. Dupas and Robinson (2013) study financial savings behavior in a developing country and find that simply providing a safe box for storing money is already quite effective for encouraging higher savings, except for the subgroup of individuals with severe present bias, who need additional social commitment. Similarly, prompting deliberation about food choice, to help resist short-run temptations, increases the effectiveness of healthy purchasing subsidies (Brownback, Imas and Kuhn, 2019).

## References

**Abrahamse, Wokje, Linda Steg, Charles Vlek, and Talib Rothengatter.** 2005. "A Review of Intervention Studies Aimed at Household Energy Conservation." *Journal of Environmental Psychology*, 25(3): 273–291.

**Agarwal, Sumit, Ximeng Fang, Lorenz Goette, Samuel Schoeb, Thorsten Staake, Verena Tiefenbeck, and Davin Wang.** 2020. "The Role of Goals in Motivating Behavior: Evidence from a Large-Scale Field Experiment on Resource Conservation." *mimeo*.

**Allcott, Hunt.** 2011. "Social Norms and Energy Conservation." *Journal of Public Economics*, 95(9-10): 1082–1095.

**Allcott, Hunt.** 2016. "Paternalism and Energy Efficiency: An Overview." *Annual Review of Economics*, 8(1): 145–176.

**Allcott, Hunt, and Todd Rogers.** 2014. "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation." *American Economic Review*, 104(10): 3003–3037.

**Andor, Mark A., and Katja M. Fels.** 2018. "Behavioral Economics and Energy Conservation – A Systematic Review of Non-price Interventions and Their Causal Effects." *Ecological Economics*, 148: 178–210.

**Andor, Mark, Andreas Gerster, Joerg Peters, and Christoph M. Schmidt.** 2020. "Social Norms and Energy Conservation Beyond the US." *Journal of Environmental Economics and Management*, 103: 102351.

**Ashraf, Nava, B. Kelsey Jack, and Emir Kamenica.** 2013. "Information and Subsidies: Complements or Substitutes?" *Journal of Economic Behavior & Organization*, 88: 133–139.

**Attari, Shahzeen Z.** 2014. "Perceptions of Water Use." *Proceedings of the National Academy of Sciences of the United States of America*, 111(14): 5129–5134.

**Attari, Shahzeen Z., Michael L. DeKay, Cliff I. Davidson, and Wändi Bruine de Bruin.** 2010. "Public Perceptions of Energy Consumption and Savings." *Proceedings of the National Academy of Sciences of the United States of America*, 107(37): 16054–16059.

**Banerjee, Abhijit, Arun Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew Jackson, Harini Kannan, Francine Loza, Anirudh Sankar, Anna Schrimpf, and Maheshwor Shrestha.** 2021. "Selecting the Most Effective Nudge: Evidence from a Large-Scale Experiment on Immunization." *Working Paper*.

**Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton.** 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *Journal of Economic Perspectives*, 31(4): 73–102.

**Brandon, Alec, John A. List, Robert D. Metcalfe, Michael K. Price, and Florian Rundhammer.** 2019. "Testing for Crowd Out in Social Nudges: Evidence from a Natural Field Experiment in the Market for Electricity." *Proceedings of the National Academy of Sciences of the United States of America*, 116(12): 5293–5298.

**Brownback, Andy, Alex Imas, and Michael Kuhn.** 2019. "Behavioral Food Subsidies." *Working Paper*.

**Camilleri, Adrian R., Richard P. Larrick, Shajuti Hossain, and Dalia Patino-Echeverri.** 2019. "Consumers Underestimate the Emissions Associated with Food but are Aided by Labels." *Nature Climate Change*, 9(1): 53–58.

**Carlsson, Fredrik, Christina Annette Gravert, Verena Kurz, and Olof Johansson-Stenman.** 2021. "The Use of Green Nudges as an Environmental Policy Instrument." *Review of Environmental Economics and Policy*, 15(2): 216–237.

**Coe, David T., and Dennis J. Snower.** 1997. "Policy Complementarities: The Case for Fundamental Labor Market Reform." *IMF Staff Papers*, 44(1).

**Cortes, Kalena E., Hans Fricke, Susanna Loeb, David S. Song, and Ben York.** 2019. "When Behavioral Barriers Are Too High or Low: How Timing Matters for Parenting Interventions." *IZA Discussion Paper No. 12416*.

**Delmas, Magali A., Miriam Fischlein, and Omar I. Asensio.** 2013. "Information Strategies and Energy Conservation Behavior: A Meta-analysis of Experimental Studies from 1975 to 2012." *Energy Policy*, 61: 729–739.

**Duflo, Esther, Pascaline Dupas, and Michael Kremer.** 2015. "Education, HIV, and Early Fertility: Experimental Evidence from Kenya." *American Economic Review*, 105(9): 2757–2797.

**Dupas, Pascaline, and Jonathan Robinson.** 2013. "Why Don't the Poor Save More? Evidence from Health Savings Experiments." *American Economic Review*, 103(4): 1138–1171.

**Dupas, Pascaline, Elise Huillery, and Juliette Seban.** 2018. "Risk information, risk salience, and adolescent sexual behavior: Experimental evidence from Cameroon." *Journal of Economic Behavior & Organization*, 145: 151–175.

**Fanghella, Valeria, Matteo Ploner, and Massimo Tavoni.** 2021. "Energy saving in a simulated environment: An online experiment of the interplay between nudges and financial incentives." *Journal of Behavioral and Experimental Economics*, 93: 101709.

**Ferraro, Paul J., and Michael K. Price.** 2013. "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment." *Review of Economics and Statistics*, 95(1): 64–73.

**Fischer, Corinna.** 2008. "Feedback on Household Electricity Consumption: A Tool for Saving Energy?" *Energy Efficiency*, 1(1): 79–104.

**Frederiks, Elisha R., Karen Stenner, and Elizabeth V. Hobman.** 2015. "Household Energy Use: Applying Behavioural Economics to Understand Consumer Decision-Making and Behaviour." *Renewable and Sustainable Energy Reviews*, 41: 1385–1394.

**Gabaix, Xavier.** 2017. "Behavioral Inattention." *NBER Working Papers 24096*.

**Gardner, Gerald T., and Paul C. Stern.** 2008. "The Short List: The Most Effective Actions U.S. Households Can Take to Curb Climate Change." *Environment and Behavior*, 50(5): 12–24.

**Gerster, Andreas, Mark Andor, and Lorenz Goette.** 2020. "Disaggregate Consumption Feedback and Energy Conservation." *CEPR Discussion Paper 14952*.

**Giaccherini, Matilde, David H. Herberich, David Jimenez-Gomez, John A. List, Giovanni Ponti, and Michael K. Price.** 2020. "Are Economics and Psychology Complements in Household Technology Diffusion? Evidence from a Natural Field Experiment." *Working Paper*.

**Hahn, Robert, Robert D. Metcalfe, David Novgorodsky, and Michael K. Price.** 2016. "The Behavioralist as Policy Designer: The Need to Test Multiple Treatment to Meet Multiple Targets." *NBER Working Paper 22886*.

**Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein.** 2014. "Learning Through Noticing: Theory and Evidence from a Field Experiment." *Quarterly Journal of Economics*, 129(3): 1311–1353.

**Holladay, J. Scott, Jacob LaRiviere, David Novgorodsky, and Michael Price.** 2019. "Prices versus nudges: What matters for search versus purchase of energy investments?" *Journal of Public Economics*, 172: 151–173.

**Imbens, Guido W., and Joshua D. Angrist.** 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467.

**Jamison, Julian C., Dean Karlan, and Jonathan Zinman.** 2014. "Financial Education and Access to Savings Accounts: Complements or Substitutes? Evidence from Ugandan Youth Clubs." *NBER Working Paper 20135.*

**Jessoe, Katrina, and David Rapson.** 2014. "Knowledge is (Less) Power: Experimental Evidence from Residential Energy Use." *American Economic Review*, 104(4): 1417–1438.

**Jessoe, Katrina, Gabriel E. Lade, Frank Loge, and Edward Spang.** 2021. "Spillovers from Behavioral Interventions: Experimental Evidence from Water and Energy Use." *Journal of the Association of Environmental and Resource Economists*, 8(2): 315–346.

**Karlin, Beth, Joanne F. Zinger, and Rebecca Ford.** 2015. "The Effects of Feedback on Energy Conservation: A Meta-analysis." *Psychological Science*, 141(6): 1205–1227.

**Kollmuss, Anja, and Julian Agyeman.** 2002. "Mind the Gap: Why Do People Act Environmentally and What Are the Barriers to Pro-Environmental Behavior?" *Environmental Education Research*, 8(3): 239–260.

**List, John A., Robert D. Metcalfe, Michael K. Price, and Florian Rundhammer.** 2017. "Harnessing Policy Complementarities to Conserve Energy: Evidence from a Natural Field Experiment." *NBER Working Paper 23355.*

**Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani.** 2019. "Inputs, Incentives, And Complementarities In Education: Experimental Evidence From Tanzania." *Quarterly Journal of Economics*, 134(3): 1627–1673.

**Myers, Erica, and Mateus Souza.** 2019. "Social Comparison Nudges Without Monetary Incentives: Evidence from Home Energy Reports." *E2e Working Paper 041.*

**Nolan, Jessica M., P. Wesley Schultz, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius.** 2008. "Normative Social Influence is Underdetected." *Personality and Social Psychology Bulletin*, 34(7): 913–923.

**Schmitt, Kathrin, Verena Tiefenbeck, Ximeng Fang, Lorenz Goette, Thorsten Staake, and Davin Wang.** 2021. "Pro-environmental spillover effects in the resource conservation domain: Evidence from a randomized controlled trial in Singapore." *mimeo.*

**Schwartz, Daniel, and George Loewenstein.** 2017. "The Chill of the Moment: Emotions and Proenvironmental Behavior." *Journal of Public Policy & Marketing*, 36(2): 255–268.

**Sherif, Raisa.** 2021. "Are Pro-environment Behaviours Substitutes or Complements? Evidence from the Field." *Max Planck Institute for Tax Law and Public Finance Working Paper 2021 – 03.*

**Tiefenbeck, Verena.** 2016. "On the Magnitude and Persistence of the Hawthorne Effect — Evidence from Four Field Studies." *4th European Conference on Behaviour and Energy Efficiency, Coimbra, Portugal.*

**Tiefenbeck, Verena, Anselma Woerner, Samuel Schoeb, Elgar Fleisch, and Thorsten Staake.** 2019. "Real-Time Feedback Promotes Energy Conservation in the Absence of Volunteer Selection Bias and Monetary Incentives." *Nature Energy*, 4: 35–41.

**Tiefenbeck, Verena, Lorenz Goette, Kathrin Degen, Vojkan Tasic, Elgar Fleisch, Rafael Lalive, and Thorsten Staake.** 2018. "Overcoming Salience Bias: How Real-Time Feedback Fosters Resource Conservation." *Management Science*, 64(3): 1458–1476.

**Tolstoy, Leo.** 2003. *Anna Karenina. (First published in Russian, 1873-1877; translation by Richard Pevear and Larissa Volokhonsky)*, London:Penguin Books.

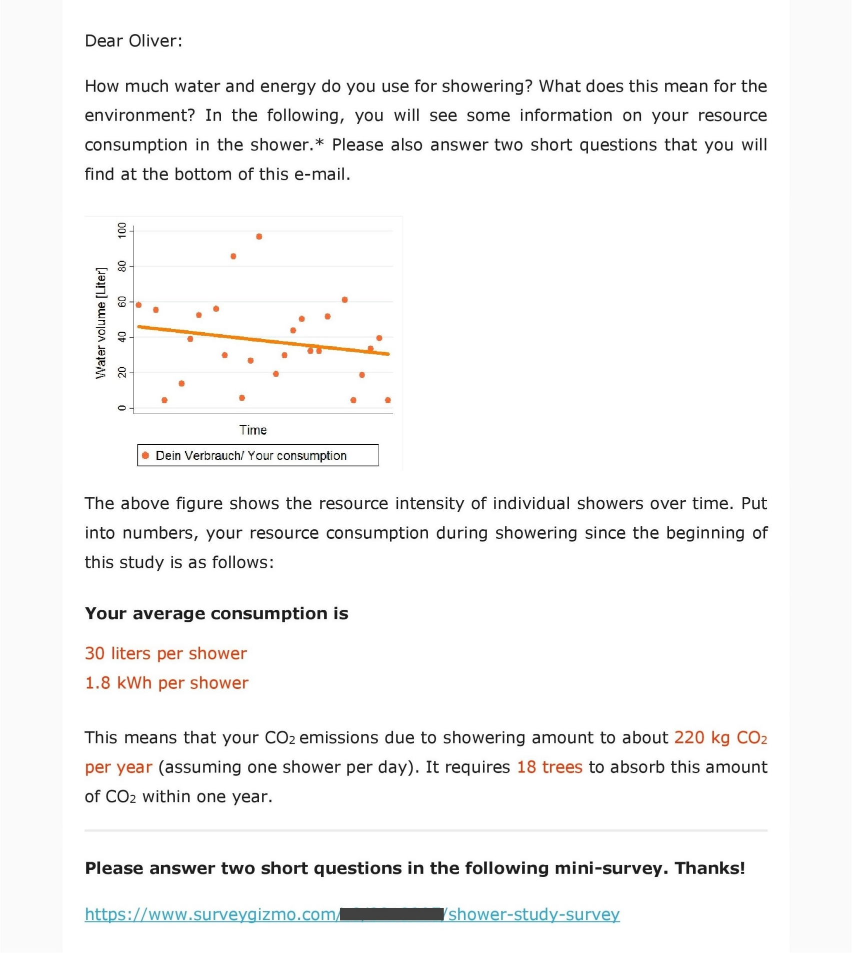
**Tonke, Sebastian.** 2019. "Imperfect Knowledge, Information Provision and Behavior: Evidence from a Field Experiment to Encourage Resource Conservation." *Working Paper*.

**Young, Alwyn.** 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *Quarterly Journal of Economics*, 134(2): 557–598.

# For Online Publication

## Appendix A Supplementary figures and tables

Figure A1: Screenshot of a typical shower energy report (for a fictitious person)



Dear Oliver:

How much water and energy do you use for showering? What does this mean for the environment? In the following, you will see some information on your resource consumption in the shower.* Please also answer two short questions that you will find at the bottom of this e-mail.

The above figure shows the resource intensity of individual showers over time. Put into numbers, your resource consumption during showering since the beginning of this study is as follows:

**Your average consumption is**

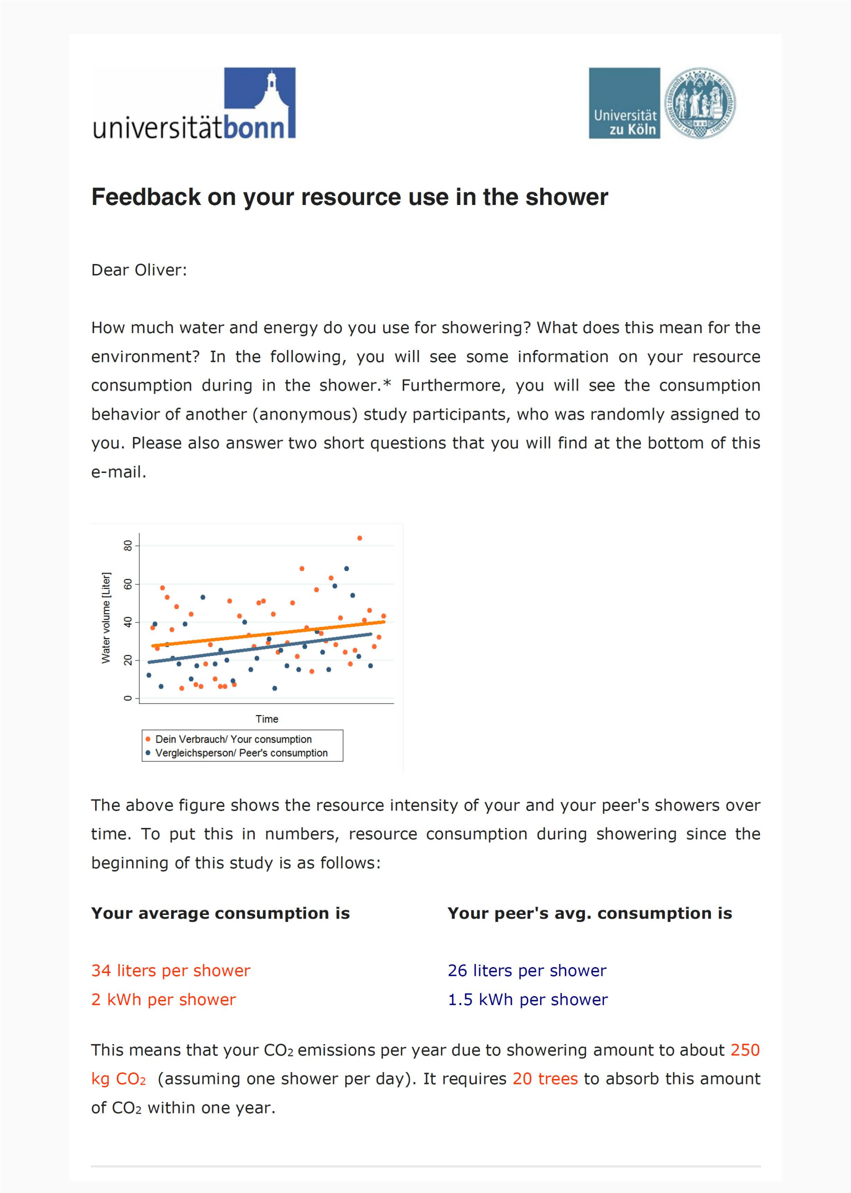
30 liters per shower
1.8 kWh per shower

This means that your $CO_2$ emissions due to showering amount to about 220 kg $CO_2$ per year (assuming one shower per day). It requires 18 trees to absorb this amount of $CO_2$ within one year.

**Please answer two short questions in the following mini-survey. Thanks!**

https://www.surveygizmo.com/[redacted]/shower-study-survey

Figure A2: Screenshot of a shower energy report with peer comparison


universitätbonn
Universität zu Köln
Water volume [Liter]
0
20
40
60
80
Time
● Dein Verbrauch/ Your consumption
● Vergleichsperson/ Peer's consumption

## Feedback on your resource use in the shower

Dear Oliver:

How much water and energy do you use for showering? What does this mean for the environment? In the following, you will see some information on your resource consumption during in the shower.* Furthermore, you will see the consumption behavior of another (anonymous) study participants, who was randomly assigned to you. Please also answer two short questions that you will find at the bottom of this e-mail.

The above figure shows the resource intensity of your and your peer's showers over time. To put this in numbers, resource consumption during showering since the beginning of this study is as follows:

| **Your average consumption is** | **Your peer's avg. consumption is** |
|---|---|
| 34 liters per shower | 26 liters per shower |
| 2 kWh per shower | 1.5 kWh per shower |

This means that your $CO_2$ emissions per year due to showering amount to about 250 kg $CO_2$ (assuming one shower per day). It requires 20 trees to absorb this amount of $CO_2$ within one year.

Figure A3: Empirical distribution of report timing

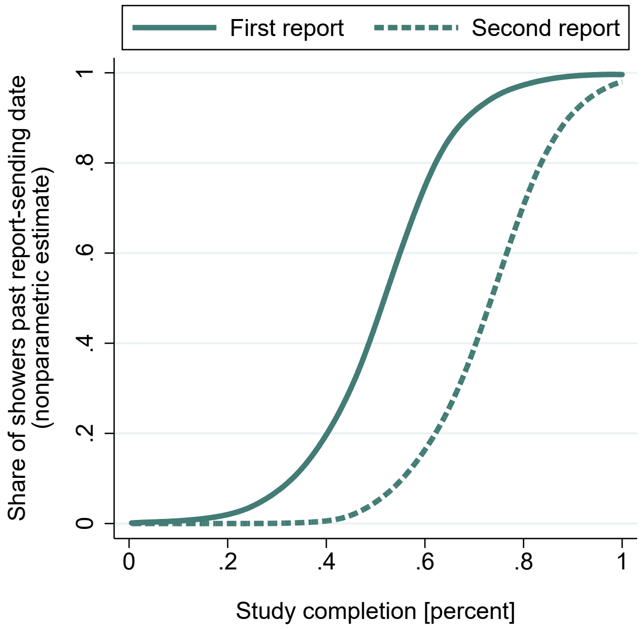


Table A1: Additional randomization checks

| | Baseline survey responses | | | | |
|---|---|---|---|---|---|
| | environmental attitude | shower comfort | 1 if female | age in years | 1 if international |
| SER group | -0.106<br>(0.165) | 0.094<br>(0.164) | -0.046<br>(0.080) | 0.757<br>(0.615) | -0.017<br>(0.075) |
| RTF group | 0.044<br>(0.167) | -0.164<br>(0.156) | -0.015<br>(0.079) | 0.872<br>(0.584) | 0.042<br>(0.077) |
| DUAL group | 0.154<br>(0.161) | 0.115<br>(0.149) | 0.117<br>(0.075) | 0.540<br>(0.583) | 0.032<br>(0.075) |
| Constant | -0.041<br>(0.118) | -0.014<br>(0.100) | 0.597<br>(0.056) | 23.351<br>(0.380) | 0.325<br>(0.054) |
| Observations | 307 | 306 | 318 | 307 | 318 |
| R-squared | 0.009 | 0.012 | 0.017 | 0.007 | 0.003 |
| F-test: *p*-value | 0.425 | 0.327 | 0.130 | 0.437 | 0.847 |

Robust standard errors in parentheses. The omitted category is the CON group.

Table A2: Comparing uploaders and non-uploaders

| | uploaders:<br>mean (sd) | non-uploaders:<br>mean (sd) | diff. in means<br>*p*-value |
|---|---|---|---|
| Energy [kWh] | 2.23<br>(1.38) | 2.20<br>(1.37) | 0.95 |
| Water volume [liter] | 38.54<br>(22.36) | 37.13<br>(20.73) | 0.87 |
| Temperature [Celsius] | 35.41<br>(3.33) | 35.94<br>(3.47) | 0.61 |
| Flow rate [liter/min] | 6.01<br>(2.34) | 5.30<br>(2.19) | 0.11 |
| Duration [min] | 6.61<br>(2.98) | 7.69<br>(4.54) | 0.10 |
| Environmental attitude | -0.04<br>(1.03) | 0.07<br>(0.93) | 0.79 |
| Shower comfort | -0.05<br>(1.05) | 0.14<br>(0.87) | 0.55 |
| 1 if female | 0.58<br>(0.49) | 0.70<br>(0.46) | 0.28 |
| Age in years | 23.93<br>(3.80) | 23.79<br>(3.99) | 0.95 |
| 1 if international | 0.31<br>(0.46) | 0.41<br>(0.49) | 0.42 |
| Observations | 228 | 90 | |

Subject characteristics before sending out shower energy reports. *p*-values adjusted for multiple hypothesis testing (Romano-Wolf procedure using $2,000$ bootstrap repetitions).

Table A3: Treatment effect dynamics

| | $Z_i = \mathbb{I}\{post\ 2nd\ report\}$ | | $Z_i = \#\ weeks\ after\ 1st\ report$ | |
|---|---|---|---|---|
| | (1)<br>Uploaders | (2)<br>LATE | (3)<br>Uploaders | (4)<br>LATE |
| Intervention | 0.179 | 0.172$^{*}$ | 0.178 | 0.171$^{*}$ |
| | (0.111) | (0.103) | (0.111) | (0.102) |
| Intervention × RTF/DUAL | -0.388$^{***}$ | -0.365$^{***}$ | -0.386$^{***}$ | -0.364$^{***}$ |
| | (0.134) | (0.125) | (0.133) | (0.125) |
| Intervention × SER | 0.027 | 0.016 | 0.029 | 0.019 |
| | (0.154) | (0.134) | (0.154) | (0.134) |
| Intervention × DUAL | 0.046 | 0.119 | 0.047 | 0.120 |
| | (0.113) | (0.108) | (0.112) | (0.108) |
| IN stage 2 | 0.139 | 0.176 | 0.065 | 0.109 |
| | (0.103) | (0.110) | (0.124) | (0.127) |
| IN stage 2 × RTF/DUAL | -0.027 | -0.053 | 0.047 | 0.019 |
| | (0.128) | (0.134) | (0.156) | (0.159) |
| IN stage 2 × SER | 0.092 | 0.048 | 0.198 | 0.174 |
| | (0.148) | (0.169) | (0.181) | (0.202) |
| IN stage 2 × DUAL | -0.068 | -0.041 | 0.030 | 0.075 |
| | (0.123) | (0.135) | (0.166) | (0.177) |
| IN stage 2 $\times Z_i$ | 0.019 | 0.022 | 0.032 | 0.029 |
| | (0.093) | (0.090) | (0.027) | (0.026) |
| IN stage 2 × RTF/DUAL $\times Z_i$ | 0.012 | 0.000 | -0.026 | -0.026 |
| | (0.123) | (0.119) | (0.037) | (0.035) |
| IN stage 2 × SER $\times Z_i$ | -0.002 | -0.010 | -0.041 | -0.051 |
| | (0.126) | (0.136) | (0.042) | (0.047) |
| IN stage 2 × DUAL $\times Z_i$ | -0.279 | -0.316 | -0.099 | -0.114$^{*}$ |
| | (0.209) | (0.215) | (0.064) | (0.067) |
| Individual fixed effects | *yes* | *yes* | *yes* | *yes* |
| Clusters | 261 | 318 | 261 | 318 |
| Observations | 14712 | 17942 | 14712 | 17942 |
| $R^2$ | 0.413 | 0.005 | 0.413 | 0.005 |

Standard errors in parentheses are clustered at the individual level. In columns (1) and (2), we exclude all non-uploaders in SER and DUAL as well as all non-uploaders in RTF and CON who did not report a technical problem. In columns (3) and (4), we use treatment assignment to SER and DUAL, respectively, interacted with the IN stage 2 indicator as instrument for receiving informative shower energy reports. The reported $R^2$ in Columns (3) and (4) is the within $R^2$.

$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

Table A4: Treatment effect heterogeneity

| | $X_i$ : *baseline energy use* | | $X_i$ : *envir. attitude* | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | linear | median$^{+}$ | linear | median$^{+}$ |
| Intervention | 0.180$^{*}$ | 0.272$^{***}$ | 0.178 | 0.243 |
| | (0.105) | (0.072) | (0.112) | (0.203) |
| Intervention × RTF/DUAL | -0.403$^{***}$ | -0.254$^{***}$ | -0.392$^{***}$ | -0.324 |
| | (0.127) | (0.096) | (0.134) | (0.222) |
| Intervention × SER | 0.012 | -0.139 | 0.003 | 0.030 |
| | (0.146) | (0.112) | (0.149) | (0.255) |
| Intervention × DUAL | 0.085 | 0.020 | 0.049 | -0.088 |
| | (0.111) | (0.087) | (0.113) | (0.151) |
| IN stage 2 | 0.148 | 0.001 | 0.166$^{*}$ | 0.140 |
| | (0.091) | (0.065) | (0.089) | (0.172) |
| IN stage 2 × RTF/DUAL | -0.014 | 0.171$^{*}$ | -0.036 | -0.032 |
| | (0.117) | (0.102) | (0.115) | (0.196) |
| IN stage 2 × SER | 0.095 | 0.267$^{**}$ | 0.074 | 0.214 |
| | (0.139) | (0.121) | (0.133) | (0.221) |
| IN stage 2 × DUAL | -0.239$^{**}$ | -0.156$^{*}$ | -0.225$^{**}$ | -0.313$^{**}$ |
| | (0.102) | (0.093) | (0.105) | (0.157) |
| Intervention $\times X_i$ | -0.016 | -0.192 | 0.031 | -0.137 |
| | (0.101) | (0.226) | (0.130) | (0.220) |
| Intervention × RTF/DUAL $\times X_i$ | -0.164 | -0.247 | -0.210 | -0.176 |
| | (0.119) | (0.266) | (0.145) | (0.269) |
| Intervention × SER $\times X_i$ | 0.109 | 0.325 | -0.172 | -0.039 |
| | (0.140) | (0.301) | (0.166) | (0.296) |
| Intervention × DUAL $\times X_i$ | 0.062 | 0.039 | 0.103 | 0.310 |
| | (0.110) | (0.215) | (0.105) | (0.232) |
| IN stage 2 $\times X_i$ | 0.056 | 0.313$^{*}$ | -0.076 | 0.056 |
| | (0.077) | (0.179) | (0.116) | (0.185) |
| IN stage 2 × RTF/DUAL $\times X_i$ | -0.094 | -0.385$^{*}$ | 0.084 | -0.002 |
| | (0.101) | (0.228) | (0.129) | (0.237) |
| IN stage 2 × SER $\times X_i$ | -0.021 | -0.368 | 0.083 | -0.363 |
| | (0.124) | (0.268) | (0.144) | (0.260) |
| IN stage 2 × DUAL $\times X_i$ | -0.097 | -0.166 | 0.024 | 0.146 |
| | (0.092) | (0.203) | (0.083) | (0.207) |
| Individual fixed effects | *yes* | *yes* | *yes* | *yes* |
| Clusters | 260 | 260 | 257 | 257 |
| Observations | 14675 | 14675 | 14501 | 14501 |
| $R^2$ | 0.413 | 0.413 | 0.414 | 0.415 |

Standard errors in parentheses are clustered at the individual level. The coefficients are obtained using the within estimator. All non-uploaders in SER and DUAL, as well as all non-uploaders in RTF and CON who did not report a technical problem, are excluded.

$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

Table A5: Estimated vs actual water use per shower

| | before study | after study | |
|---|---|---|---|
| | | ITT | TOT |
| Actual volume | 0.271 | 0.175 | 0.186 |
| | (0.263) | (0.139) | (0.145) |
| Actual volume × RTF | 0.025 | 0.742*** | 0.835*** |
| | (0.376) | (0.199) | (0.179) |
| Actual volume × SER | -0.465 | 0.289* | 0.381** |
| | (0.292) | (0.174) | (0.169) |
| Actual volume × DUAL | -0.074 | 0.520*** | 0.517** |
| | (0.299) | (0.182) | (0.230) |
| RTF group | -0.131 | 1.694 | 3.162 |
| | (6.777) | (3.234) | (3.183) |
| SER group | -7.001 | -4.578 | -5.200* |
| | (5.813) | (3.181) | (3.029) |
| DUAL group | -5.182 | 1.655 | 1.588 |
| | (5.851) | (3.136) | (3.826) |
| Constant | 43.436*** | 39.507*** | 39.610*** |
| | (4.590) | (2.429) | (2.542) |
| Observations | 267 | 296 | 251 |
| $R^2$ | 0.030 | 0.378 | 0.440 |

Robust standard errors in parentheses. Actual volume is recentered around 40 liters.

$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

Table A6: Estimated versus actual water use: relative estimation error

| | before study | after study | |
|---|---|---|---|
| | | ITT | TOT |
| RTF group | 0.075 | -0.283*** | -0.296*** |
| | (0.201) | (0.073) | (0.075) |
| SER group | 0.008 | -0.172** | -0.281*** |
| | (0.175) | (0.080) | (0.072) |
| DUAL group | -0.055 | -0.214** | -0.270*** |
| | (0.178) | (0.085) | (0.076) |
| Constant | 0.927*** | 0.577*** | 0.583*** |
| | (0.136) | (0.061) | (0.064) |
| Observations | 302 | 296 | 251 |
| $R^2$ | 0.002 | 0.050 | 0.101 |

Robust standard errors in parentheses.

$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

Table A7: Response to mini-surveys attached to reports

| | Survey response rate | | | |
|---|---|---|---|---|
| | (1)<br>first report | (2)<br>second report | (3)<br>any report | (4)<br>estimation error [%p] |
| RTF group | -1.05 | 0.53 | -2.48 | -30.69 |
| | (5.35) | (6.57) | (4.90) | (7.62) |
| SER group | -7.85 | -16.76 | -7.18 | -38.74 |
| | (6.39) | (7.91) | (5.81) | (7.54) |
| DUAL group | 0.58 | -26.17 | -0.44 | -27.70 |
| | (5.54) | (8.14) | (5.00) | (8.57) |
| Constant | 88.89 | 80.56 | 91.67 | 48.93 |
| | (3.73) | (4.70) | (3.28) | (7.14) |
| *p*-value for SER = DUAL | 0.203 | 0.308 | 0.270 | 0.039 |
| Observations | 261 | 261 | 261 | 231 |
| R-squared | 0.009 | 0.061 | 0.008 | 0.139 |

Robust standard errors in parentheses.

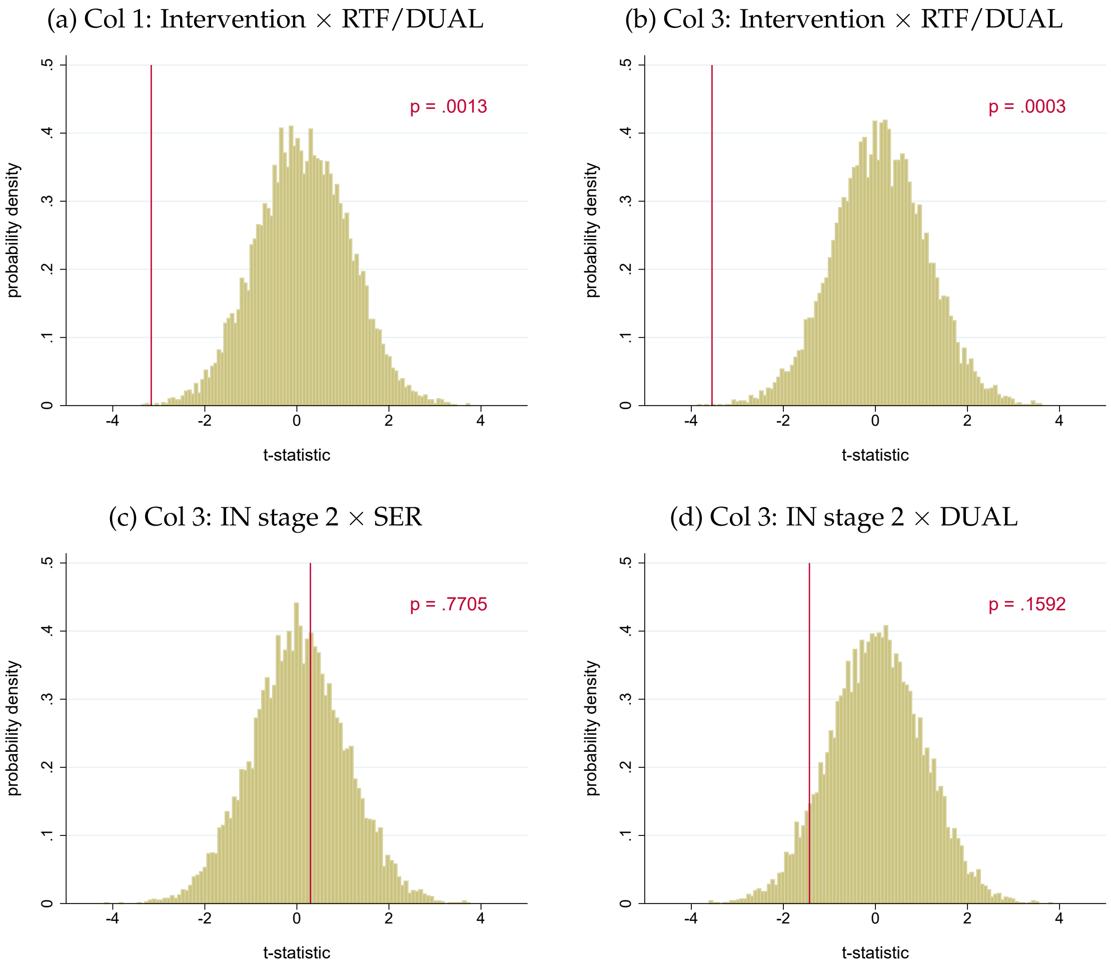
Table A8: Margins of behavioral adjustment

| | Duration in seconds | | | Temperature in °C | | | Flow rate in liter/min | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1)<br>ITT | (2)<br>Uploaders | (3)<br>LATE | (4)<br>ITT | (5)<br>Uploaders | (6)<br>LATE | (7)<br>ITT | (8)<br>Uploaders | (9)<br>LATE |
| Intervention | 8.40 | 8.52 | 8.40 | -0.00 | 0.01 | -0.00 | 0.25$^{**}$ | 0.25$^{**}$ | 0.25$^{**}$ |
| | (11.02) | (11.76) | (11.02) | (0.34) | (0.37) | (0.34) | (0.11) | (0.12) | (0.11) |
| Intervention × RTF/DUAL | -38.41$^{**}$ | -39.32$^{**}$ | -38.41$^{**}$ | -0.74 | -0.88$^{*}$ | -0.74 | -0.17 | -0.19 | -0.17 |
| | (16.31) | (17.17) | (16.31) | (0.46) | (0.49) | (0.46) | (0.17) | (0.18) | (0.17) |
| Intervention × SER | 13.37 | 11.16 | 11.50 | -0.54 | -0.63 | -0.53 | -0.08 | -0.12 | -0.07 |
| | (16.35) | (17.60) | (16.19) | (0.42) | (0.46) | (0.41) | (0.15) | (0.18) | (0.15) |
| Intervention × DUAL | 6.35 | 7.73 | 6.32 | -0.23 | 0.19 | -0.25 | -0.01 | -0.20 | -0.04 |
| | (16.92) | (17.49) | (16.55) | (0.40) | (0.41) | (0.38) | (0.18) | (0.19) | (0.17) |
| IN stage 2 | 24.69 | 12.99 | 24.69 | 0.39 | 0.46 | 0.39 | 0.26$^{**}$ | 0.24$^{**}$ | 0.26$^{**}$ |
| | (17.75) | (9.81) | (17.75) | (0.31) | (0.34) | (0.31) | (0.11) | (0.12) | (0.11) |
| IN stage 2 × RTF/DUAL | -32.30 | -19.75 | -32.30 | 0.28 | 0.31 | 0.28 | 0.21 | 0.22 | 0.21 |
| | (19.74) | (13.33) | (19.74) | (0.40) | (0.42) | (0.40) | (0.19) | (0.20) | (0.19) |
| IN stage 2 × SER | -30.85 | -12.45 | -38.01 | 0.18 | 0.21 | 0.22 | 0.16 | 0.11 | 0.20 |
| | (21.83) | (17.21) | (27.21) | (0.37) | (0.40) | (0.46) | (0.17) | (0.17) | (0.21) |
| IN stage 2 × DUAL | -0.49 | 0.71 | -0.62 | -0.21 | -0.41 | -0.26 | -0.34 | -0.40 | -0.43 |
| | (12.70) | (14.22) | (16.01) | (0.35) | (0.33) | (0.44) | (0.22) | (0.25) | (0.28) |
| Observations | 17942 | 14712 | 17942 | 17942 | 14712 | 17942 | 17942 | 14712 | 17942 |
| $R^2$ | 0.383 | 0.361 | 0.001 | 0.310 | 0.323 | 0.003 | 0.751 | 0.763 | 0.016 |

Standard errors in parentheses (clustered on subject level)

$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

Figure A4: Randomization inference for coefficients of interest in Table 3



(a) Col 1: Intervention × RTF/DUAL

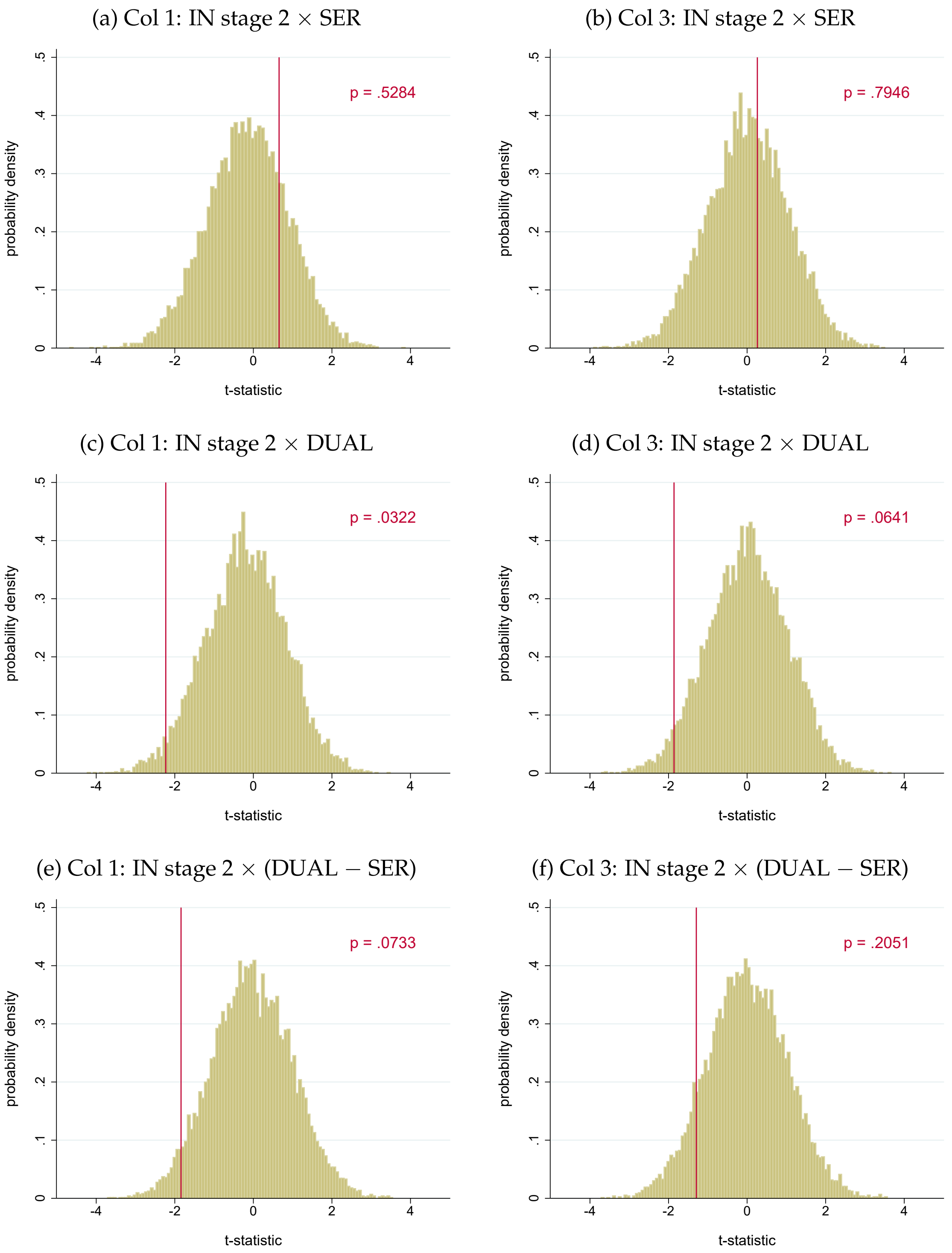
(b) Col 3: Intervention × RTF/DUAL

(c) Col 3: IN stage 2 × SER

(d) Col 3: IN stage 2 × DUAL

*Notes.* Distribution of estimated t-statistics based on 10,000 permutation samples. For each permutation, treatment assignment into CON, SER, RTF, or DUAL was randomly relabeled, holding constant the actual number of individuals in each treatment group. The red vertical line represents the t-value for the true treatment labels. Permutation-based *p*-values are shown in the top right corner.

Figure A6: Randomization inference for coefficients of interest in Table 4


(a) Col 1: IN stage 2 × SER

(b) Col 3: IN stage 2 × SER

(c) Col 1: IN stage 2 × DUAL

(d) Col 3: IN stage 2 × DUAL

(e) Col 1: IN stage 2 × (DUAL − SER)

(f) Col 3: IN stage 2 × (DUAL − SER)

*Notes.* Distribution of estimated t-statistics based on 10,000 permutation samples. For each permutation, treatment assignment into CON, SER, RTF, or DUAL was randomly relabeled, holding constant the actual number of individuals in each treatment group. The red vertical line represents the t-value for the true treatment labels. Permutation-based *p*-values are shown in the top right corner.

## Appendix B Randomization protocol

At the beginning of the study, we randomly assigned subjects into groups that receive or do not receive real-time feedback. Each smart meter was programmed as either treatment or control device. Treatment device started displaying real-time feedback from the eleventh shower onwards, whereas control devices only ever showed the current water temperature. When distributing the smart meters to subjects, we alternated between treatment and control devices after each apartment. Thus, treatment and control devices are by construction balanced within dorms.

We assigned subjects into groups with or without shower energy report shortly before we intended to sent out the reports. We used the data that subjects uploaded through the smartphone app to rank them from lowest to highest average water use per shower, split by whether they receive real-time feedback or not. Then, we formed pairs between subjects adjacent to each other in rank and assigned shower energy reports to only one member of a pair based on a virtual coin flip. This ensures that the distribution of resource consumption levels remain balanced across experimental conditions. Subjects who had not uploaded any data at that point in time were assigned to a group randomly without prior ranking.

The second shower energy report further contained a social comparison component with a random and anonymous peer. This peer was assigned to subjects in the following way: (1) we used uploaded data prior to the second report to rank subjects again by their average water use per shower; (2) we then selected three potential peers for each subject, a subject who was somewhat above him/her in rank, a subject who was somewhat below him/her in rank, and a directly adjacent subject; (3) we then chose one of these three candidates randomly with equal probabilities; (4) subjects who had not uploaded any data received a random peer from the pool of subjects who had uploaded data. This procedure ensured that the direction of peer comparison was orthogonal to subjects' resource use level.

## Appendix C Data cleaning procedures

A number of data cleaning steps are performed before running the empirical analyses. In principle, we have access to the smart meter data from two sources: (1) uploads by subjects themselves using the smartphone app, and (2) the data that we read out manually after retrieving the devices. For the large majority of devices, the two sources gave us identical data. In the cases where it differed, we always opted to use the information we read out manually.

We drop the very first data point of each participant, as they usually started with a test run to check if the device was working. Following Tiefenbeck et al. (2018), we further drop any water extraction with volume below 4.5 liters (in total $2,942$ extractions), as these are unlikely to be actual showers but rather minor extractions for other purposes such as cleaning. We further remove 37 extreme outlier points, defined as energy use and water use for that shower being more than 4.5 times the subject-specific interquartile range away from the closest quartile. We are particularly strict in only excluding the most unplausible data points here. Conventionally, 1.5 or 3 times the interquartile range (IQR) are used as criterion for outliers. For a normal distribution, 4.5 times the IQR away from the nearest quartile corresponds to 6.745 standard deviation away from the mean.

We further exclude 1 device with erratic data, as evidenced by huge intra-device variance (the largest for all devices) and some outrageous data points with water volumes of up to above 500 liters for a single shower. In 8 cases, the device's temperature sensor broke at some point, and we impute missing information with the average temperature of showers taken while the sensor was still intact. For some devices, we detected an error through which decimal places of the flow rate are shifted such that the stored number is actually ten times the actual flow rate. We corrected these manually for showers with flow rates that are about ten times the flow rate of other showers stored on the device.

## Appendix D Timing of showers

As the smart meter itself has no global time counter and only stores the chronological order of water extractions, we make use of smartphone app information to put a time stamp on each observation. In particular, we need to determine whether a shower took place before or after we sent out the shower energy reports, so whether it is in intervention stage 2. The app provides us with information on the date and time of each data upload by subjects. This allows us construct time windows in which a shower observation has plausibly happened. Firstly, a shower must have been taken by the time data was uploaded via the app, so this gives us the upper bound. Secondly, it must have been taken place after the previous data upload, because otherwise it would have been uploaded by then; this gives us the lower bound. To be able to determine the timing relatively reliably around the crucial time period, in which we sent out shower energy reports, we sent several upload reminders to all participants. Whenever it was not unambiguously clear, which shower was the first that took place after a shower energy report, we assigned the switching point implied by constant shower frequency. For example, if one upload was 1 day before the shower energy report and the next upload 1 day after, and there were 2 showers in the window, we assumed that the first shower was before and the second shower after the report.

A complication arising from non-uploaders is that we do not know the timing of showers by these participants, because the shower meter itself only stores the order of showers but not the time and date. We can only infer the earliest and latest possible date of each shower based on when it was uploaded to the smartphone app. Therefore, whenever we want to include non-uploaders in our analyses, we need to impute the timing of showers in one way or another, in particular whether it took place before or after a shower energy report.

We use a pragmatic imputation approach based on the assumption that, given the stage of study completion, i.e. which fraction of the number of total recorded showers have been completed, showers by uploaders and non-uploaders have the same probability of having taken place after the first/second shower energy report. Formally, we assume that for each stage of study completion $\pi$,

$$Pr\left(IN_{it}^{s2} = 1 | \pi, non\text{-}uploader\right) = Pr\left(IN_{it}^{s2} = 1 | \pi, uploader\right).$$

To operationalize this approach, we estimate the distribution of uploaders' report timing over study completion non-parametrically, so $\widehat{Pr}\left(IN_{it}^{s2} = 1 \mid \pi, uploader\right)$, and, instead of the indicator $IN_{\pi}^{s2}$ for intervention stage 2, we define

$$\widehat{IN}_{s}^{s2} = \widehat{Pr}\left(IN_{it}^{s2} = 1 \mid \pi_{it}^{s} = 1, uploader\right)$$

as probabilistic indicator for every shower of non-uploaders in study completion stage $\pi$. In other words, the regressor $\widehat{IN}_{\pi}^{s2}$ is the probability that a particular shower by a non-uploader took place after the first shower energy report. In all our regressions, we actually use the indicator

$$\widetilde{IN}_{it}^{s2} = \begin{cases} IN_{it}^{s2} & \textit{if uploader} \\ \widehat{Pr}\left(IN_{it}^{s2} = 1 \mid \pi, uploader\right) & \textit{if non-uploader}\,. \end{cases} \tag{14}$$

## Appendix E Supplementary Survey

We conducted a supplementary survey in a new sample of students in November and December 2019, about three years after the original experiment took place. The purpose of the survey was two-fold. First, we wanted to collect evidence that people tend to underestimate the environmental impact of showering without additional information. Second, we wanted to provide a manipulation check for our shower energy report intervention, testing whether the additional information on energy use and $CO_2$ emissions due to showering can plausibly induce stronger conservation efforts. The survey was conducted among residents of exactly the same student dorms in Bonn and Cologne in which the original study took place. Thus, the surveyee pool is comparable to the subject pool of the original experiment. In total, 329 students participated in the supplementary survey. Due to the high fluctuation rate of residents in student dorms, only 4 out of the 329 surveyees had also participated in the original experiment in 2016/17.

We first elicited students' prior beliefs about the amount of water used and $CO_2$ emitted per shower, as well as how confident they are about their response on a 10-point scale. As reference, we told surveyees that one hour of room lighting causes about 10 grams of

$CO_2$ and that one hour of watching TV causes about 30 grans of $CO_2$. Furthermore, we asked students about their intention to take shorter showers on a 10-point Likert scale (we normalize this to mean 0 and standard deviation 1 for all analyses). After the first round of questions, we randomly presented one fact sheet (out of three) to each surveyee, mimicing the basic informational content of our original interventions. The "CON sheet" only contained information on average water temperature in the shower, the "RTF sheet" also included the average water use per shower, and the "SER sheet" further added information on energy use and $CO_2$ emissions. The exact wording was as follows. All fact sheets started with this text:

*"Did you know that a few years ago, a study was conducted in this dorm, as well as other dorms in Cologne and Bonn? The study has shown that the average water temperature when taking a shower is about 37 degrees Celsius."*

While the CON sheet ended here, the RTF sheet added the sentence *"... A typical shower uses around 40 liters of water."*. The SER sheet provided even more information by adding the following sentences: *"... A typical shower uses around 40 liters of water and 2.4 kWh of energy. This means that, on average, a person's emissions due to daily showering amount to almost 300 kg CO2 per year (800 grams per shower). It requires about 24 trees to absorb this amount of CO2."*. After surveyees had finished reading their respective fact sheet, we elicited posterior beliefs and attitudes by asking them the same questions again that they answered before receiving additional information. Surveyees were then paid 5 Euros for their participation in the survey, although 11 students refused to accept any remuneration.

Prior to receiving the fact sheets, surveyees estimated on average that they use 40.4 liters of water per shower (standard error of the mean = 6.36), causing emissions of 91.3 grams of $CO_2$ (s.e.m. = 15.03). While the estimate for water used per shower is roughly accurate on average, surveyees grossly underestimate the amount of $CO_2$ emitted by a factor of 8 to 9. However, subjects are also very uncertain about their estimates. On a scale from 1 (very uncertain) to 10 (very certain), the average surveyee places him-/herself at 4.24 for water use and 3.71 for $CO_2$ emissions.

Table A9 shows how surveyee change their beliefs and intentions after being provided with additional information through the fact sheets. Neither the RTF nor the SER survey induces statistically significant changes in surveyees' average estimates for water use per shower compared to the CON sheet, although surveyees in these groups become much more confident about their answer. In contrast, only the SER fact sheet has a strong impact on surveyees beliefs about $CO_2$ emissions. As surveyees severely underestimated the carbon intensity of showering in baseline, the SER fact sheet had an extreme debiasing effect compared to the CON and RTF fact sheets. This experimentally-induced belief update about environmental impacts is further associated with a sizeable increase in self-stated intentions to take shorter showers. Compared to surveyees receiving the RTF sheet, conservation intentions of surveyees receiving the SER sheet increased by 0.24 standard deviations ($p = 0.003$). In contrast, the RTF sheet did not increase intentions

Table A9: Supplementary survey — change in beliefs and intentions after fact sheet

| | Water use per shower | | $CO_2$ emissions | | |
|---|---|---|---|---|---|
| | (1)<br>Estimate | (2)<br>Confidence | (3)<br>Estimate | (4)<br>Confidence | (5)<br>Intention |
| RTF fact sheet | -12.274 | $2.148^{***}$ | 28.774 | $0.358^{*}$ | 0.060 |
| | (10.146) | (0.279) | (21.587) | (0.208) | (0.065) |
| SER fact sheet | -22.909 | $2.561^{***}$ | $484.941^{***}$ | $2.023^{***}$ | $0.304^{***}$ |
| | (16.813) | (0.258) | (37.599) | (0.264) | (0.076) |
| Constant | 14.203 | 0.118 | -15.274 | $0.335^{**}$ | $0.088^{**}$ |
| | (9.663) | (0.161) | (19.023) | (0.138) | (0.042) |
| $p$-value for RTF = SER | 0.451 | 0.175 | 0.000 | 0.000 | 0.003 |
| Baseline mean | 40.428 | 4.239 | 91.335 | 3.711 | 0.000 |
| Observations | 328 | 328 | 329 | 329 | 329 |
| $R^2$ | 0.008 | 0.222 | 0.476 | 0.185 | 0.054 |

Robust standard errors in parentheses. The omitted category is the CON fact sheet group. Column (1) and (2) exclude one subjects who did not give a baseline estimate for water use. The intention measure used for column (5) is normalized to mean 0 and standard deviation 1.

significantly compared to the CON sheet ($p = 0.359$). Overall, these results suggests that people tend to severely underestimate the environmental impact of showering, and that information provision about energy and carbon intensity can induce subjects to increase their conservation efforts.

## Appendix F More on Other Potential Mechanisms

### F.1 Hawthorne or cueing effects

Given that we observe energy and water use in a relatively private and sensitive activity, showering, subjects' behavior may have been distorted by Hawthorne effects. We attempt to hold this constant by equipping every participant with a functioning smart shower meter, so to the degree that subjects in the control group respond to the sheer presence of a shower meter (with temperature feedback), we would in fact underestimate our conservation effects. To explain our empirical findings, Hawthorne effects would thus need to additionally interact with the intervention regimes. As the conservation effect in the RTF group (compared to the CON group) is quantitatively large and remains stable over the entire 3-months study duration, it seems unlikely that it is driven by differential Hawthorne effects. However, the shower energy reports may have made it more salient again to participants that they were part of a study, or alternatively, the reports may have simply served as a general cue or reminder to pay more attention to conservation efforts in the shower. Note that we sent out placebo emails instead of infor-

Table A10: Change in self-reported attitudes (baseline vs. post-intervention survey)

| | *shower comfort* | | *environmental attitude* | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | ITT | TOT | ITT | TOT |
| RTF group | 0.042 | 0.047 | -0.340*** | -0.345*** |
| | (0.117) | (0.119) | (0.117) | (0.119) |
| SER group | 0.085 | 0.090 | -0.277** | -0.253* |
| | (0.134) | (0.136) | (0.133) | (0.145) |
| DUAL group | -0.097 | -0.011 | -0.225* | -0.239* |
| | (0.138) | (0.150) | (0.129) | (0.144) |
| Constant | 0.026 | 0.030 | 0.139 | 0.143 |
| | (0.086) | (0.088) | (0.094) | (0.095) |
| F-test: $p$-value | 0.641 | 0.896 | 0.034 | 0.039 |
| Observations | 300 | 255 | 304 | 257 |
| $R^2$ | 0.007 | 0.003 | 0.027 | 0.031 |

Robust standard errors in parentheses.
$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

mative shower energy reports to the RTF and CON groups precisely to limit such types of confounders. Furthermore, we find that, if anything, the effect of shower energy reports (in the DUAL group) tends to become stronger over time instead of weaker, and the exercise in Figure 9 using different complier definitions for LATE estimation also suggests that it is the actual content of shower energy reports that matters. While we have no way to directly rule out Hawthorne or cueing effects, we are therefore confident that they do drive our empirical results.

## F.2 Environmental attitude and consumption value of showering

Another alternative way in which two interventions could develop complementarities is through some sort of motivational spillover effect, in which the combined intervention convinced subjects to generally care more for the environment, or somehow made showering less pleasurable to them. Our interventions presented all information in a neutral and factual way, and we specifically refrained from including any normative element. Nevertheless, to check if this could confound our results, we again analyse subjects' survey responses before and after the study. The outcome variable of interest is the change in environmental attitude index or shower comfort index, respectively. All indices are normalized by subtracting the pre-intervention mean and dividing by the pre-intervention standard deviation.

The first two columns in Table A10 show difference-in-differences estimates for the effect of treatments on subjective shower comfort from baseline to endline survey. Both in the ITT (column 1) and in the TOT (column 2) regressions for subjective shower comfort, we find no significant differences across experimental condition, and all point estimates are virtually zero. Hence, at least based on self-reported measures, our interventions do not seem to have diminished the consumption benefits of showering, which is also relevant for welfare considerations.

The other two columns in Table A10 show the difference-in-differences estimates for impacts on environmental attitude, with ITT estimates in column (3) and TOT estimates in column (4). Surprisingly, we find that subjects in the treated groups become *less* pro-environmental relative to the control group based on their survey responses. The magnitude of this decrease ranges from 22% to 35% of a (pre-study) standard deviation, which is not exactly quantitatively large, but also not negligible. We can only speculate about what is happening here. At face value, it may seem that feedback makes people less motivated to act pro-environmentally. Of course, we only have self-reported measures and cannot be certain of the underlying latent variable that they proxy for. But as we seem to proxy *self-perceived* inclination to act pro-environmentally rather than the actual extent of pro-environmental behavior, one possible interpretation could be that feedback provision curbs the capacity for distorted self-image formation, because people become aware of their intention-action gaps. We caution from overinterpreting the result here, as we did not have any ex ante hypothesis along these lines. Still, we can tentatively conclude that the conservation effects we observe are unlikely due to generally increased pro-environmental motivation.